Beatriz Wilges

# UM MODELO PARA ORGANIZAÇÃO DE DOCUMENTOS NO CONTEXTO DA MEMÓRIA ORGANIZACIONAL

Tese submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Grau de Doutor em Engenharia e Gestão do Conhecimento.
Orientador: Prof. Dr. Rogério Cid Bastos.
Coorientadora: Prof.ª Drª. Lia Caetano Bastos.

Florianópolis
2014

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Wilges, Beatriz
Um modelo para organização de documentos no contexto da memória organizacional / Beatriz Wilges ; orientador, Rogério Cid Bastos ; coorientadora, Lia Caetano Bastos. - Florianópolis, SC, 2014.
125 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Inclui referências

1. Engenharia e Gestão do Conhecimento. 2. Modelo de organização de documentos. 3. Memória organizacional. 4. Classificação de texto em múltiplas categorias. 5. Modelagem fuzzy. I. Cid Bastos, Rogério . II. Caetano Bastos, Lia . III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Beatriz Wilges

**UM MODELO PARA ORGANIZAÇÃO DE DOCUMENTOS NO CONTEXTO DA MEMÓRIA ORGANIZACIONAL**

Esta Tese foi julgada adequada para obtenção do Título de "Doutor em Engenharia e Gestão do Conhecimento", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 03 de julho de 2014.

Prof. Gregório Varvakis, Dr.
Coordenador do Curso

Prof. Rogério Cid Bastos, Dr.
Orientador

**Banca Examinadora:**

Prof. Júlio Cesar Nievola, Dr.
PUCPR

Prof. Fernando Alvaro Ostuni Gauthier, Dr.
UFSC

Prof. Jorge Luiz Silva Hermenegildo, Dr.
IFSC

Prof. Mario Antonio Ribeiro Dantas, Dr.
UFSC

Prof.ª Silvia Modesto Nassar, Dr.ª
UFSC

## AGRADECIMENTOS

Ao meu orientador, Rogério Cid Bastos, pelo apoio e dedicação. Seus conselhos foram fundamentais para a realização deste trabalho.

À minha professora e orientadora no mestrado, Silvia Modesto Nassar. Sua cooperação, revisão e apoio foram essenciais para que esta pesquisa se tornasse realidade.

Aos professores Júlio Nievola, Jorge Hermenegildo, Fernando Gauthier e Mario Dantas, por aceitarem o convite para a banca de qualificação e defesa final. Suas contribuições na etapa de qualificação foram decisivas para a qualidade deste trabalho.

Agradeço com muito carinho ao Gustavo Mateus, pela ajuda, compreensão e paciência durante todos os momentos que passamos juntos. Seu apoio na execução dos testes e verificações foi fundamental.

Aos meus familiares, pelos momentos com os quais não pude compartilhar.

Agradeço, também, a todos que contribuíram com o desenvolvimento desta tese: aos colegas, amigos e professores. Obrigada pelo incentivo e colaboração no andamento e finalização deste trabalho.

Com o aumento da complexidade,
declarações precisas perdem o significado e
declarações significativas perdem a precisão.

Lotfi Zadeh

## RESUMO

Gerenciar e estruturar um conjunto de documentos em uma organização pode otimizar os processos de gestão, contribuindo para o seu desempenho e sucesso. Sabe-se que, apesar de haver iniciativas de gestão do conhecimento (GC), a quantidade de informações heterogêneas muitas vezes inviabiliza uma gestão produtiva. A memória organizacional (MO) fornece acesso, persistência e recuperação de dados. Assim, esta pesquisa se concentrou na definição da estrutura de um modelo de organização de documentos para a MO, o qual é apoiado por um método desenvolvido para a classificação dos documentos em múltiplas categorias com lógica *fuzzy*. Para avaliação deste modelo, considerou-se a estrutura de uma organização de tecnologia da informação (TI) com um conjunto de 17 categorias. Os resultados agregam valor para a organização porque permitem tratar um conjunto de informações espalhadas em diversos documentos, refinando o espaço de busca e recuperando a informação de interesse para os indivíduos que nela atuam. Além disso, o trabalho individual migra para um nível coletivo, porque se pode tratar informações de interesse comum aos grupos dentro da organização.



**Palavras-chave**: Modelo de organização de documentos. Memória organizacional. Classificação de texto em múltiplas categorias. Modelagem *fuzzy*.

## ABSTRACT

The task of managing and structuring a set of documents within an organization can optimize the management process, contributing to it's performance and success. It is well known that despite the efforts of knowledge management (KM), the amount of heterogeneous information often prevents a productive management. The organizational memory (OM) provides access, persistence and retrieval of data. Thus, this research has focused on the definition of the structure of a document organization model in the context of an organizational memory, which is supported by a methodology developed for the classification of documents into multiple categories with *fuzzy* logic. The structure of an information technology (IT) organization with a set of 17 categories was considered to evaluate this model. The results add value to the organization because they allow to treat a set of information spread over different documents, refining the search space and retrieving relevant information to individuals working in it. Additionally, individual work migrates to a collective level, because it can handle information of common interest to the groups within the organization.



**Keywords**: Document organization model. Organizational memory. Multiple categories text classification. *Fuzzy* modeling.

# LISTA DE FIGURAS

# LISTA DE QUADROS

# LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| $A_{PT}^{C}$ | Confiança (*accuracy*) do texto analisado em relação à categoria |
| AD | Árvores de decisão |
| BD | Banco de dados (*Database*) |
| BR | Relevância binária (*binary relevance*) |
| c | Categoria |
| CADT | Categorização automática de documentos de texto |
| $C_A$ | Confiança alta |
| $C_B$ | Confiança baixa |
| $C_M$ | Confiança média |
| DM | *Data mining* |
| EC | Engenharia do conhecimento |
| $f_{(w_i,c_j)}$ | Frequência das palavras no BD para uma categoria |
| GC | Gestão do conhecimento |
| KDD | *Knowledge discovery in database* |
| k-NN | *k-nearest neighbors* |
| LC | *Label combination* |
| LF | Lógica *fuzzy* |
| LP | *Label powerset* |
| LSA | *Latent semantic analysis* |
| LSI | *Latent semantic indexing* |
| MC | Mídias e conhecimento |
| ML-kNN | *Multi-label k-Nearest Neighbors* |
| MO | Memória organizacional |
| PIC | Processos intensivos em conhecimento |
| PLN | Processamento de linguagem natural |
| PPT | *Pruned problem transformation* |
| PS | *Pruned set* |
| $P_S$ | Texto pertence |
| PT | Texto analisado (*parsed text*) |
| RAkEL | *Random k-labelsets* |

| | |
|---|---|
| $RD_{w_i}^{c_j}$ | Grau de relevância (*relevance degree*) |
| RI | Recuperação da informação |
| RPC | *Ranking by pairwise comparision* |
| $S_A$ | Similaridade alta |
| $S_B$ | Similaridade baixa |
| $S_{PT}^{c}$ | Similaridade de um texto em relação à categoria |
| $S_M$ | Similaridade média |
| SVM | *Support vector machines* |
| T | Documento de texto |
| TI | Tecnologia da informação |
| tf-idf | *Term frequency–inverse document frequency* |
| w | Palavra (*word*) |

**SUMÁRIO**

# 1 INTRODUÇÃO

Um modelo para organização de documentos pode ser um fator diferencial e essencial para viabilizar a gestão dos diferentes conteúdos que envolvem o âmbito organizacional. Segundo Neumann (2013), a palavra organização se refere ao modo pelo qual se organiza um sistema para cumprir certas funções. Uma organização é um sistema complexo, em que o todo tem propriedades e capacidades que as partes isoladamente não têm. No entanto, desenvolver um modelo para estruturar documentos para uma organização não é uma tarefa fácil, especialmente quando se trata dos chamados processos intensivos em conhecimento (PIC). Os PICs compreendem desde atividades baseadas na aquisição até atividades de compartilhamento, armazenamento e reutilização do conhecimento.

De acordo com Wei et al. (2014), as organizações têm participado cada vez mais dos ambientes da internet para realizar transações comerciais, reunindo informações de *marketing* e inteligência competitiva a partir de várias fontes *on-line*, o que facilita o compartilhamento da informação e do conhecimento dentro e fora dos seus limites. Essas aplicações de comércio eletrônico e gestão do conhecimento geram e mantêm uma quantidade enorme de documentos textuais em repositórios organizacionais. Para facilitar o acesso posterior a esses documentos, o uso de categorias para gerenciar o seu crescente volume muitas vezes ocorre em níveis organizacionais.

De acordo com Toledo et al. (2011), no mercado global o conhecimento é considerado um ativo que tem um valor econômico para a organização, além de ser um recurso estratégico utilizado para aumentar a produtividade e oferecer estabilidade em ambientes competitivos e dinâmicos. A importância do conhecimento está relacionada com seu acesso direto, sua persistência ao longo do tempo e a possibilidade de recuperação, quando necessário.

Uma forma de prover o acesso, a persistência e a recuperação de dados em uma organização é a utilização de uma memória organizacional (MO). Segundo Sasieta, Beppler e Pacheco (2011), o conceito de MO pode ser entendido como a habilidade da organização em salvar, reter e fazer uso de informações do passado nas atividades atuais. É um elemento-chave que permite que ela aprenda com os erros e acertos do passado. Mas fazer uso dessas informações é uma atividade complexa, pois elas estão dispersas em diferentes lugares e são heterogêneas.

A Engenharia do Conhecimento (EC) fornece um conjunto de ferramentas que pode ser usado para adquirir informações e conhecimento a partir de diversas fontes heterogêneas. Além disso, essas informações podem ser organizadas de modo a permitir correlacionamentos que forneçam informações desconhecidas, até então, para a organização.

Dessa forma, esta pesquisa tem a seguinte questão: Como prover um modelo para organizar documentos no contexto de uma organização?

De acordo com Ale et al. (2008), existem muitas iniciativas de gestão do conhecimento (GC) implementadas nas organizações, mas a maioria desses esforços muitas vezes não consegue gerenciar a heterogeneidade das suas fontes de conhecimento.

Assim, esta pesquisa apresenta um modelo para a organização de conteúdo apoiado por um processo de classificação de documentos. A classificação de documentos permite que o conhecimento seja estruturado e organizado e que, se necessário, sejam adicionadas novas fontes que possibilitem que cada área de atuação da organização administre de forma autônoma o seu próprio repositório de conhecimento e o compartilhe com outras áreas organizacionais.

A classificação de documentos ou a categorização automática de documentos de texto (CADT), no ambiente de uma organização, é fundamental para a gestão dos artefatos que envolvem tanto as informações como o conhecimento organizacional. A importância das pesquisas relacionadas com a CADT vem ao encontro da abordagem de gestão do conhecimento (GC), porque, no contexto de uma organização, fatores que permitem a estruturação da informação propiciam não só a qualidade desta, mas também o ganho de tempo na busca de uma informação específica.

Segundo Supyuenyong e Swierczek (2011), os benefícios da GC são reconhecidos principalmente nas grandes organizações, mas pequenas e médias empresas também podem obter resultados positivos. Em suas pesquisas em diferentes organizações, os resultados mostram que organizar, reter e utilizar o conhecimento pode melhorar o desempenho individual, o desempenho do produto e o desempenho global da organização.

Isso significa, entre outras coisas, que as pessoas dentro da organização têm mais acesso ao conhecimento porque a informação está disponível de modo organizado. Assim, os sistemas de CADT passam a ter um papel importante no processo de organização de documentos, auxiliando nos processos de busca, como a recuperação de informação.

De acordo com Ren e Sohrab (2013), o processo de recuperação de documentos em bases de dados devidamente rotuladas gera classificadores mais eficazes, além de reduzir o espaço de busca de um grande volume de informações.

## 1.1 OBJETIVO

O objetivo geral desta pesquisa é propor um modelo para organização de documentos por meio de um processo de classificação no contexto da memória organizacional (MO).

### 1.1.1 Objetivos específicos

- Desenvolver um método para classificação dos documentos considerando múltiplas categorias;
- identificar um conjunto de regras para a classificação dos documentos por meio de um processo de extração de conhecimento em base de dados (*Knowledge Discovery in Databases* - KDD);
- estabelecer procedimentos que tratem a incerteza dos parâmetros utilizados na classificação.

## 1.2 RELEVÂNCIA DA PESQUISA

A importância desta pesquisa está fundamentada na contribuição oferecida por meio da estruturação e organização dos documentos dentro de uma MO. De acordo com Manne e Fatima (2012), dados categorizados são fáceis de serem pesquisados se estiverem bem organizados.

O valor agregado desse modelo, para uma organização, está essencialmente focado na melhoria dos processos de gestão dos artefatos de conhecimento organizacional, permitindo que documentos de interesse sejam acessados exclusivamente pelas pessoas interessadas. Além disso, o modelo organizacional proposto relaciona tarefas intensivas em conhecimento como a aquisição, compartilhamento, armazenamento e reutilização do conhecimento. Ou seja, existe a articulação entre diferentes abordagens e técnicas de engenharia do conhecimento (EC), gestão do conhecimento (GC) e até mesmo de

mídias e conhecimento (MC). Assim, sua natureza interdisciplinar permite um aprimoramento do conhecimento na medida em que as diferentes áreas são articuladas para construção de um modelo de organização de documentos.

Entre as diferentes técnicas de EC trabalhadas para obter esse modelo, são abordados aspectos de extração de conhecimento em base de dados (*Knowledge Discovery in Databases* - KDD) para reconhecimento de padrões, utilizando algoritmos de mineração de dados para classificação tanto preditiva quanto descritiva, além da técnica de lógica *fuzzy* para construção de uma modelagem difusa.

A proposta do modelo de organização de documentos, em função do processo de classificação, foi definida para apoiar, também, a melhoria da gestão do conhecimento (GC) dentro da organização. Segundo Silva (2008), a GC, além de auxiliar a organização em seu desempenho nas ações estratégicas, também auxilia o processo de inteligência competitiva organizacional subsidiando a geração de ideias, solucionando problemas e melhorando a tomada de decisão.

Apesar de haver pesquisas envolvendo processos de categorização de documentos na literatura, um método para classificação de textos, em múltiplas categorias, no contexto de uma modelagem difusa, ainda não foi devidamente explorado. O resultado de um método como este poderia indicar graus de pertinência de um documento em várias categorias de interesse.

De acordo com Antonie e Zaiane (2002) e Kim et al. (2009), um procedimento de classificação de textos considerado bom para um grande conjunto de documentos deve classificar com uma precisão aceitável, ter regras de classificação legíveis para possíveis ajustes e processar em um tempo razoável. De acordo com a literatura, alguns procedimentos são mais precisos, outros mais interpretáveis e outros realizam em um tempo mais aceitável, porém nenhuma das pesquisas estudadas combinou todas essas propriedades.

Além disso, a maioria das pesquisas trata o problema de classificação de documentos de texto em um único rótulo, no qual um documento só pode pertencer a uma categoria (JIANG; TSAI; LEE, 2012). Em pesquisas onde a classificação do texto envolve múltiplos rótulos, é possível que um documento pertença a mais de uma categoria.

## 1.3 CONTRIBUIÇÕES DA PESQUISA

A representação das atividades empresariais intensivas em conhecimento é necessária para se chegar a uma compreensão adequada do que deve ser apoiado na gestão dos negócios de uma organização. Na pesquisa de França, Santoro e Baião (2012) é proposta uma ontologia para definição e caracterização dos PICs, cujo objetivo é a consolidação dos conceitos necessários para uma definição precisa de um PIC. No entanto de acordo com França et al. (2012), é difícil encontrar um método que aborde a totalidade ou, pelo menos, a maior parte dessas características na representação dos seus processos, principalmente por causa da falta de métodos de análise adequados. Além disso, a colaboração entre os participantes e a manutenção de uma memória organizacional (MO) podem ser determinantes no valor do conhecimento agregado para a organização.

Assim, um modelo de organização de documentos que atue diretamente no contexto da MO pode trazer avanços e potencializar o seu uso. Além de propor esse modelo, esta pesquisa pretende contribuir com o desenvolvimento e especificação de um método para classificação de documentos, considerando uma abordagem em múltiplas categorias com lógica *fuzzy*.

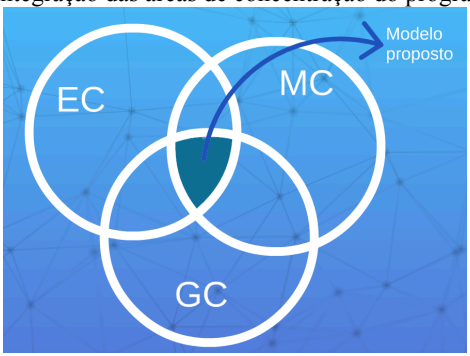
Dessa forma, o modelo proposto é capaz de tratar, com a lógica *fuzzy*, a sobreposição de categorias que existe na classificação de um documento, retornando o grau de pertinência do documento em categorias de interesse dentro da organização. Assim, o uso da MO é potencializado porque permite identificar documentos relevantes para um indivíduo, cuja existência ele desconhecia por estar alocado em outro setor da organização.

## 1.4 ADERÊNCIA E INTERDISCIPLINARIDADE

No contexto desta pesquisa, um método de classificação e categorização de documentos é essencial para desenvolver o modelo de organização de documentos sobre a MO. Entende-se que a relevância científica desta pesquisa está em sua natureza interdisciplinar capaz de integrar a área de Gestão do Conhecimento (GC) no que tange ao desempenho organizacional, à engenharia do conhecimento (EC), às atividades intensivas em conhecimento e à área de mídias e

conhecimento (MC) na manipulação e disponibilização de documentos digitais (Figura 1).

Figura 1 – Integração das áreas de concentração do programa de EGC




Fonte: Elaborada pela autora

O paradigma desta pesquisa se enquadra na filosofia funcionalista. O paradigma funcionalista possui uma orientação pragmática, em que a compreensão deve ser posta em termos de conhecimentos gerais que depois devem ser colocados em prática. Por isso é uma abordagem orientada por problemas que se preocupa em prover soluções práticas.

Este trabalho é também caracterizado como uma pesquisa tecnológica, porque procura meios para construir e implementar o que propõe. Ou seja, este trabalho tem uma visão mais aplicada e incorpora em sua proposta a preocupação com a utilização prática do artefato final produzido.

Assim, a concepção deste trabalho perpassa o paradigma funcionalista no sentido de buscar soluções objetivas e também é caracterizado como uma pesquisa tecnológica, porque está centrado em um conjunto de atividades direcionadas para produzir e utilizar um artefato. Para isso, diferentes técnicas são empregadas na definição do método de classificação do modelo proposto. O cenário interdisciplinar, nesta pesquisa, busca a combinação dessas diferentes abordagens porque espera uma solução que potencialize a arquitetura da MO por meio desse modelo de organização de documentos.

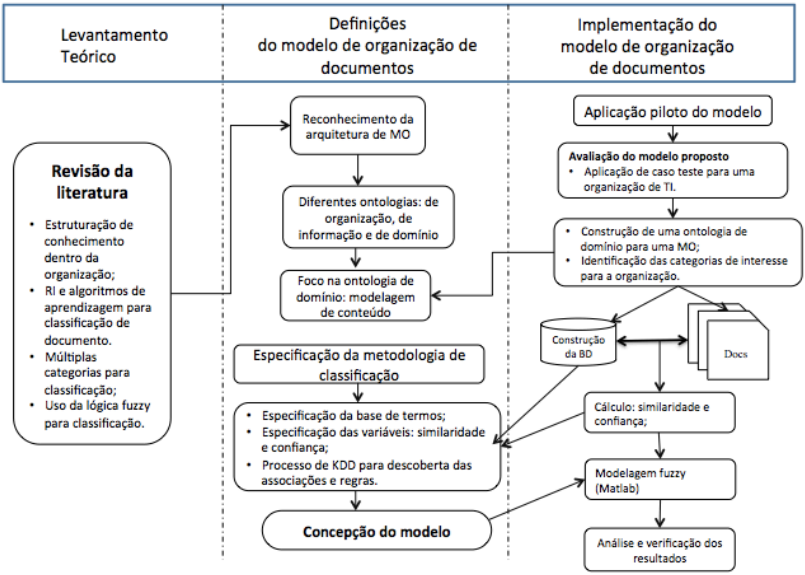
## 1.5 PROCEDIMENTOS METODOLÓGICOS

O desenho desse modelo de organização de documentos no cenário de uma memória organizacional pode ser uma atividade complexa, por envolver uma grande quantidade de informações e conhecimentos que necessitam ser categorizados ou mapeados dentro dos interesses da organização. Segundo Nonaka, Toyama e Nagata (2000), a organização não é simplesmente uma máquina de processamento de informações, mas uma entidade que cria conhecimento por meio da ação e interação.

A classificação de documentos, nesta pesquisa, engloba mais de uma categoria e admite o conceito de uma abordagem difusa para a classificação. Ou seja, entende-se que o processo de classificação, além de envolver *n* categorias, considera a imprecisão em seu processo. Assim, o resultado final desse método de classificação permitirá atribuir graus de pertencimento do documento para cada categoria.

Esta pesquisa planeja e desenvolve um modelo de organização de documentos por meio da construção de um método de classificação. A seguir apresenta-se, em síntese, o fluxograma das atividades metodológicas previstas na construção do modelo (Figura 2). Além do levantamento teórico, é necessário definir o escopo e os parâmetros bem como a avaliação do modelo proposto.

Figura 2 – Síntese dos procedimentos metodológicos




Fonte: Elaborada pela autora

No centro do fluxograma são relatadas a especificação e a descrição do modelo proposto. Nessa parte, é definido o escopo de atuação desse modelo no âmbito da memória organizacional (MO). Ou seja, a partir de uma arquitetura de MO reconhecida, é definido o campo de atuação desse modelo. Além disso, todos os detalhes que envolvem a descrição do desenvolvimento do método de classificação são descritos nessa etapa.

A especificação da base de termos por categoria considerou, de acordo com o SestatNet (2014), um tamanho mínimo da amostra calculado pela seguinte fórmula:

$$n' = \left(\frac{z}{e_a}\right)^2 . P(1-P) \qquad \text{(Equação 1)}$$

Onde:

$n'$: amostra calculada;

$z$: valor normal padronizado em relação ao nível de confiança adotado (z=1,96 para 95% de nível de confiança);

$e_a$: margem de erro amostral tolerável;

$P$: valor do percentual estimado. Adota-se $P$=0,5 dado que, neste caso, a variabilidade é máxima.

O valor $n'$ pode ser corrigido considerando o tamanho ($N$) da população de estudo resultando em uma amostra de *n* elementos:

$$n = \frac{n'}{\left(1 + \frac{n'}{N}\right)} \qquad \text{(Equação 2)}$$

Onde:

$n$: amostra corrigida;
$N$: população.

Também nessa etapa são definidas as variáveis e são gerados dados para uma base de conhecimento com essas variáveis. Essa base é construída para permitir que um processo de descoberta de conhecimento encontre associações entre as variáveis para a concepção de regras para uma modelagem difusa.

Assim, parte dessas atividades envolve um processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* - KDD), especificamente em um conjunto de aprendizagem dividido em treinamento e teste. O processamento sobre esse conjunto de treinamento é realizado por algoritmos de classificação. A avaliação do modelo gerado nessa fase é feita através do cálculo da acurácia e do erro. Tanto a acurácia, quanto o erro do modelo consideram os resultados das correspondências entre a classe real e a classe prevista. Dessa forma, seus valores são obtidos da seguinte forma:

$$A = \frac{VP + VN}{n} x\ 100\% \qquad \text{(Equação 3)}$$

$$E = \frac{FP + FN}{n} x 100\% \qquad \text{(Equação 4)}$$

Onde:
$VP$ = verdadeiros positivos;
$VN$ = verdadeiros negativos;
$FN$ = falsos negativos;
$FP$ = falsos positivos;
$n = VP + VN + FP + FN$.

A etapa de implementação do modelo de organização de documentos é a aplicação prática dessa proposta (terceira parte). Assim, é necessária uma base de dados de uma MO para formalizar um conjunto com diversos termos para todas as categorias identificadas como de interesse para organização. A partir dessa base de dados são realizados os cálculos de similaridade e confiança, que são as entradas para a modelagem de classificação *fuzzy*.

Espera-se que esse modelo seja efetivo e arranje de forma adequada a estrutura dos documentos analisados, sendo capaz de exibir os graus de pertinência para cada uma das categorias definidas, as quais necessitam ser identificadas a partir dos interesses da organização.

Este trabalho está organizado da seguinte forma: no segundo capítulo apresenta-se a revisão da literatura, que apoia o método de classificação de documentos proposto para a construção desse modelo. O terceiro capítulo apresenta a concepção e definição do modelo de organização de documentos. O quarto capítulo descreve a implementação do modelo proposto, o quinto capítulo apresenta a avaliação do modelo e o sexto capítulo, as considerações e recomendações deste trabalho.

# 2 REVISÃO DA LITERATURA

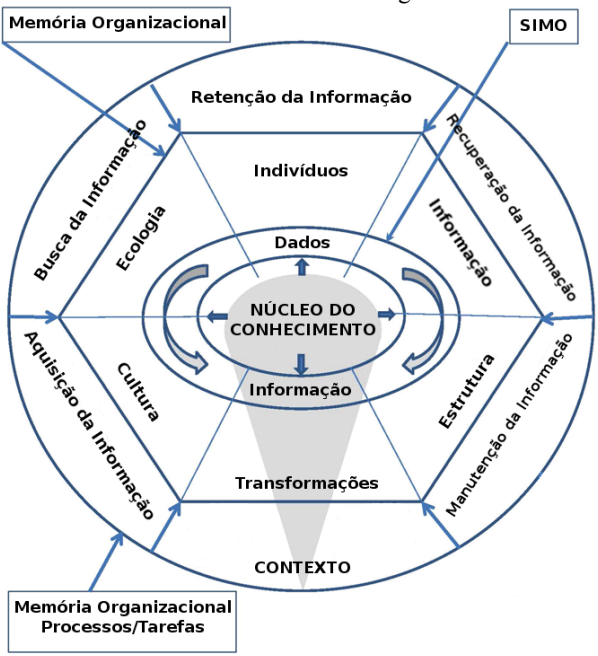
A revisão da literatura trata dos diferentes conceitos que são a chave para a construção do modelo de organização de documentos. Nessa análise abordam-se as definições de memória organizacional (MO), com um metamodelo representado por ontologias de uma MO. Apresentam-se neste capítulo o processo de categorização automática de documentos de texto (CADT), as abordagens de recuperação da informação (RI), os diferentes algoritmos de aprendizagem utilizados para classificação de documentos, os processos de classificação com lógica difusa, a classificação em múltiplas categorias e uma descrição sobre o processo de mineração de dados para descoberta de conhecimento.

## 2.1 MEMÓRIA ORGANIZACIONAL

De acordo com Jackson (2008), a MO é o conhecimento de como fazer as coisas, a forma de abordar os problemas e suas questões. Segundo Argote (2013), a MO se preocupa com a reutilização e compartilhamento do conhecimento para utilizá-lo nas atividades atuais, melhorando assim sua eficácia organizacional. O conceito de memória organizacional está intrinsecamente vinculado ao conceito de aprendizagem organizacional. Para Argote e Miron-Spektor (2011), embora os pesquisadores tenham definido aprendizagem organizacional de diferentes formas, no núcleo da maioria das definições considera-se que esta é uma mudança na organização que ocorre quando ela adquire experiência, ou seja, quando o conhecimento é o resultado do aprendizado.

Segundo Dow, Hackbarth e Wong (2013), a memória corporativa ou organizacional no núcleo de uma organização de aprendizagem deve apoiar o compartilhamento e a reutilização de conhecimento individual e empresarial, e as lições aprendidas devem estar dispostas em torno dessa MO. Os serviços de gestão de conhecimento inteligentes fornecem ativamente ao usuário que trabalha em uma tarefa operacional o conhecimento intensivo com toda a informação necessária e útil para a realização de sua tarefa (Figura 3).

Figura 3 – A MO nas atividades básicas de gestão de conhecimento




Fonte: Adaptada de Dow, Hackbarth e Wong (2013)

A Figura 3 apresenta no centro o núcleo do conhecimento, composto pelo ciclo dos dados e informações, gerenciado pelo Sistema de Informação da Memória Organizacional (SIMO). A partir do núcleo, tem-se a representação da MO composta pelos indivíduos, ecologia, cultura, transformações, estrutura e informações. Ao redor da MO estão os PICs: aquisição, manutenção, recuperação, retenção e busca da informação.

O modelo utilizado para organizar documentos está inserido no contexto de uma organização, mais especificamente na estrutura da Memória Organizacional (MO). Assim, esse modelo potencializa os processos intensivos em conhecimento (PIC), porque reorganiza as informações armazenadas em repositórios a partir de um conjunto de categorias bases.

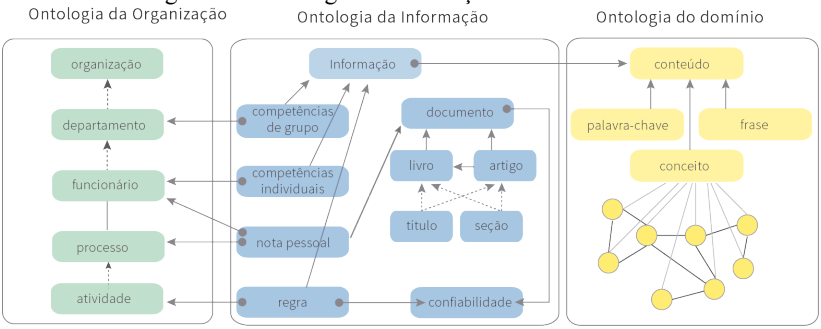
Ainda segundo Dow, Hackbarth e Wong (2013), a função principal de uma MO é aumentar a competitividade da organização, melhorando a forma de gerenciar seu conhecimento. Para atingir esse objetivo, os esforços devem se concentrar na preservação do conhecimento que se baseia em grande parte na explicação do

conhecimento tácito. O conhecimento tácito é apoiado por sistemas especialistas, sistemas de informação temáticos, bancos de dados de melhores práticas e arquivos de lições aprendidas. Segundo Toledo et al. (2011), a aquisição do conhecimento e sua forma de representação são cruciais para a memória organizacional.

### 2.1.1 Metamodelo baseado em ontologias

Nesta pesquisa, analisou-se a representação da MO por meio de um metamodelo baseado em três ontologias que descrevem os itens de informação e conhecimento. Essa abordagem foi proposta por Abecker et al. (1998), que apresenta esse metamodelo e descreve as diferentes origens da informação com suas respectivas estruturas, formatos e propriedades de acesso (Figura 4).

Figura 4 – Ontologias de descrição de conhecimento




Fonte: Abecker et al. (1998)

A ontologia da organização é utilizada para descrever os contextos que envolvem os itens de conhecimento. Os conceitos fornecidos pela ontologia da informação contêm termos e atributos genéricos que podem ser aplicados a todos os tipos de informação. Basicamente, essa ontologia compreende todos os aspectos de informação e conhecimento, os quais não são específicos ao conteúdo. A ontologia de informação também fornece vínculos à ontologia de domínio, que é utilizada para descrever o conteúdo, e vínculos à ontologia da organização.

O modelo de organização desta pesquisa é aplicado especificamente na apresentação desse metamodelo de MO, no contexto da ontologia de domínio. Assim, um método para classificação dos

documentos é aplicado sobre o conjunto de documentos considerado na ontologia de domínio, observando seu conteúdo, palavras-chave, frases e conceitos.

A próxima seção descreverá os aspectos do processo de categorização de documentos de textos. Esse processo tem como base a aplicação da ontologia de domínio proposta no modelo de organização de documentos da MO.

## 2.2 CATEGORIZAÇÃO DE DOCUMENTOS DE TEXTO

Segundo Abdul-Rahman et al. (2013), com o crescente número de documentos é difícil para os usuários procurar, encontrar, gerenciar e organizar as informações rapidamente. Assim, um processo capaz de realizar a atribuição de documentos a um conjunto de categorias predefinidas, agrupando documentos semanticamente relacionados, como a técnica de CADT, pode apoiar a construção de um modelo de organização de documentos.

Encontrar informações relevantes, e em um tempo aceitável, em documentos, é importante para muitas aplicações, como: organização desses documentos, busca automática e acesso à informação. Além disso, os procedimentos de CADT são úteis em áreas como: gestão do relacionamento com clientes (COUSSEMENT; VAN DEN POEL, 2008), filtragem de *spam* em *e-mail* (ZHOU; YAO; LUO, 2010) e classificação de páginas *web* (QI; DAVISON, 2009).

Segundo Azam e Yao (2012), a CADT desempenha um papel importante em aplicações nas quais a informação deve ser filtrada, monitorada, personalizada, categorizada, organizada ou pesquisada. Assim, em vez de selecionar um documento entre milhares, pode-se analisar apenas aqueles pertencentes às categorias de maior interesse.

A CADT, quanto utilizada na recuperação de informações, procura reduzir o espaço de busca, facilitando o acesso à informação e permitindo a indexação por assunto. Além disso, ela é de grande importância para as organizações porque permite o tratamento e a classificação dos documentos por meio de informações que são consideradas relevantes. Isso possibilita um processo de organização e triagem de documentos de interesse dentro da organização. Na Figura 5 apresenta-se esquematicamente o processo de categorização de documentos de texto.

Figura 5 – Processo de categorização de textos


Fonte: Elaborada pela autora

De acordo com Hao, Ying e Longyuan (2009), a CADT baseada em aprendizagem de máquina é composta por duas etapas: a aprendizagem e a categorização. Antes de iniciar o processo de aprendizagem é necessário realizar o pré-processamento dos dados. Esta etapa é responsável por processar os dados do documento bruto para que sejam representados de uma forma adequada para aplicar o algoritmo de classificação, tanto para a etapa de aprendizagem/treinamento como para a categorização.

O pré-processamento é uma etapa-chave fundamental para sistemas de categorização de texto. Uysal e Gunal (2014) realizaram um trabalho com o objetivo de analisar o impacto da etapa de pré-processamento na classificação de textos em aspectos como: precisão da classificação, domínio do texto, linguagem do texto e redução de dimensões. Análises experimentais nos conjuntos de dados revelam que a escolha de combinações adequadas para as tarefas de pré-processamento pode proporcionar uma melhoria significativa na precisão da classificação.

Na fase de aprendizagem, estuda-se a classificação do conhecimento por meio de um conjunto de treinamento de textos e estabelece-se um classificador. O estudo do modelo procede quando pode ser identificada cada classe na amostra de treinamento, e os métodos de aprendizagem podem ser dados na forma de regras de classificação, árvores de decisão ou fórmulas matemáticas. Em seguida, é necessário testar e avaliar a taxa de precisão do modelo de aprendizagem, prever e comparar com classes conhecidas. Na fase de categorização, o texto de entrada é classificado na categoria mais provável, de acordo com o classificador (HAO; YING; LONGYUAN, 2009).

Segundo Russel e Norvig (2009), quanto mais complexo for o processo de categorização, mais difícil e demorado será o tratamento da informação, obrigando a combinação de técnicas de processamento de

linguagem natural, recuperação da informação e métodos de análise de dados qualitativos.

Classificadores de texto têm sido propostos na literatura, utilizando técnicas de aprendizagem de máquina e abordagens de recuperação da informação (RI). Geralmente os algoritmos de categorização são baseados em resultados obtidos de aprendizagem de máquinas, porém muitas propostas partem da representação formal dos documentos segundo as abordagens definidas pela área de RI. Os algoritmos de aprendizagem de máquina diferem na abordagem adotada: árvores de decisão, Naïve-Bayes, regras de indução, redes neurais, os vizinhos mais próximos (*k-Nearest Neighbors* – k-NN) e, também, máquinas de vetores de suporte (*Support Vector Machines* – SVM).

Embora muitas abordagens tenham sido propostas, a CADT ainda é uma das principais áreas de pesquisa, principalmente porque a eficácia dos classificadores de texto automáticos não é impecável e ainda pode ser melhorada (ANTONIE; ZAIANE, 2002; IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005; KUMAR; SUGANITHI, 2013).

## 2.3 RECUPERAÇÃO DA INFORMAÇÃO

As três abordagens para recuperação da informação (RI) mais difundidas são: modelo booleano, modelo probabilístico e modelo vetorial. Trabalhos relacionados com a representação de documentos e linguagem de consulta são as questões mais importantes em qualquer pesquisa de RI.

Segundo Manning, Raghavan e Schütze (2008), o modelo booleano expressa uma consulta por meio de seus conectivos lógicos AND, OR ou NOT. Nesse modelo um documento é considerado somente como relevante ou não relevante para uma consulta, ou seja, não existe um resultado parcial e não há informações suficientes para permitir a ordenação dos resultados por relevância. Apesar de ser simples sua programação, esse modelo pode apresentar resultados nulos na saída ou apresentar diversos resultados sem considerar uma ordenação.

A segunda abordagem é o modelo probabilístico, o qual propõe o processo de recuperação como um problema de estimativa da probabilidade de que a representação de um documento corresponda ou satisfaça a representação de uma consulta. Ou seja, dada uma consulta de um usuário, existe um conjunto que contém exatamente os documentos relevantes e outro conjunto de documentos não relevantes.

Segundo Jones, Walker e Robertson (2000), dada uma consulta $q$ e um documento $d_j$ em uma coleção, o modelo probabilístico tenta estimar a probabilidade de um usuário considerar relevante o documento $d_j$. Esse modelo assume que essa probabilidade de relevância depende apenas da representação da consulta e da representação do documento.

Ainda segundo esses autores, o modelo probabilístico assume que existe um subconjunto de documentos que o usuário prefere como resposta para a consulta $q$. O conjunto de documentos ideais é representado por $R$ e deve maximizar a probabilidade de relevância para o usuário. Documentos no conjunto $R$ são rotulados como relevantes para a consulta $q$. Documentos que não estão nesse conjunto são considerados não relevantes para $q$ e rotulados como $\bar{R}$, o complemento de $R$.

Na pesquisa de Colace, Santo e Greco (2014) apresentou-se um método de classificação de texto em uma única categoria utilizando uma abordagem probabilística e foi utilizado um vetor de características estruturadas, compostas por pares de palavras ponderadas. O diferencial dessa pesquisa se baseou no tratamento de um pequeno conjunto de documentos categorizados para desenvolver o modelo proposto.

Na terceira abordagem, o modelo vetorial, um documento é representado em um espaço *T*-dimensional como um vetor:

$$\vec{d} = (w_{d,1}, w_{d,2}, \dots, w_{d,T}) \qquad \text{(Equação 5)}$$

Onde *T* representa o número de termos distintos da coleção e $w_{d,i}$ representa o peso do termo *i* no documento *d*. Assim cada documento é representado como um vetor de termos e cada termo tem um valor associado que indica o seu grau de importância em um determinado documento.

Da mesma forma, os termos de uma consulta também são associados a um peso. O vetor para uma consulta $\vec{q}$ é definido como $\vec{q} = (w_{q,1}, w_{q,2}, \dots, w_{tq,T})$, onde $w_{q,1} \geq 0$, *T* é o número total de termos na consulta e $w_{q,i}$ representa o peso do termo *i* para a consulta $q$.

Ainda no modelo vetorial, é utilizado um cálculo para o grau de similaridade entre o documento e a consulta denominado de "similaridade dos cossenos". Essa correlação utiliza o cosseno do ângulo entre dois vetores $\vec{d}$ e $\vec{q}$ e é definida da seguinte forma:

$$sim(\vec{d}, \vec{q}) = \frac{\sum_{i=1}^{T} w_{d,i} \times w_{q,i}}{\sqrt{\sum_{i=1}^{T} w_{d,i}^2} \times \sqrt{\sum_{i=1}^{T} w_{q,i}^2}} \qquad \text{(Equação 6)}$$

Onde:

$\sum_{i=1}^{T} w_{d,i} \times w_{q,i}$, representa o somatório da multiplicação do peso do termo *i* no documento *d* pelo peso do termo *i* para a consulta *q*;

$\sqrt{\sum_{i=1}^{T} w_{d,i}^2}$ representa a norma do vetor documento, calculada pelo somatório do quadrado do peso do termo para o documento;

$\sqrt{\sum_{i=1}^{T} w_{q,i}^2}$, representa a norma do vetor consulta, calculada pelo somatório do quadrado do peso do termo para a consulta.

Dessa forma, o modelo vetorial permite ordenar os documentos de acordo com o grau de similaridade de cada um com a consulta realizada pelo usuário. Assim, um documento pode ser recuperado mesmo que satisfaça a consulta apenas parcialmente (BAEZA-YATES; RIBEIRO-NETO, 2011; MANNING; RAGHAVAN; SCHÜTZE, 2008).

O peso dos termos nos documentos pode ser calculado de várias formas, dependendo das características da base de termos e do tipo de recuperação ou categorização que são realizadas. Uma das formas mais conhecidas de calcular o peso de um termo em um documento é tentar balancear o número de ocorrências do termo no documento com o número de documentos onde o termo aparece. Essa equação é definida como tf-idf (*term frequency–inverse document frequency*) e é descrita da seguinte forma:

$$w_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} \times \log \frac{N}{n_i} \qquad \text{(Equação 7)}$$

Onde:

$w_{i,j}$ = peso do termo $i$ no documento $j$;

$N$ = número de documentos da coleção;

$n_i$ = número de documentos onde a palavra $i$ aparece;

$f_{i,j}$ = frequência normalizada da palavra $i$ no documento $j$; $max_z f_{z,j}$ = frequência da palavra $z$ mais encontrada no documento $j$.

Dessa forma, segundo Zadrożny e Kacprzyk (2006), o peso é proporcional à frequência do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece. Trata-se de uma boa forma de contabilizar os termos, determinando um peso maior a um termo se ele é um bom discriminante. Ou seja, se um termo aparece em uma quantidade maior de documentos, ele deve ser mais significativo para a classificação em uma determinada categoria.

Muitas pesquisas sobre RI focalizam estruturas que possam tratar melhor a semântica das informações. Nessas estruturas entram técnicas de processamento de linguagem natural (PLN) como a Indexação com Semântica Latente (*Latent Semantic Indexing* – LSI), também denominada Análise com Semântica Latente (*Latent Semantic Analysis* – LSA).

Nas pesquisas de Antonie e Zaiane (2002), desenvolveu-se uma abordagem para a CADT que utiliza mineração de regras de associação na fase de aprendizado. Além disso, o modelo desenvolvido por esses autores permite que as saídas possam ser lidas, entendidas e modificadas manualmente. Apesar de apresentarem resultados interessantes em seus experimentos, eles afirmam que aplicar técnicas de seleção de características, como Análise com Semântica Latente (LSA), poderia melhorar os resultados obtidos. Essa técnica de seleção de característica escolhe um subconjunto de características relevantes para a construção do modelo.

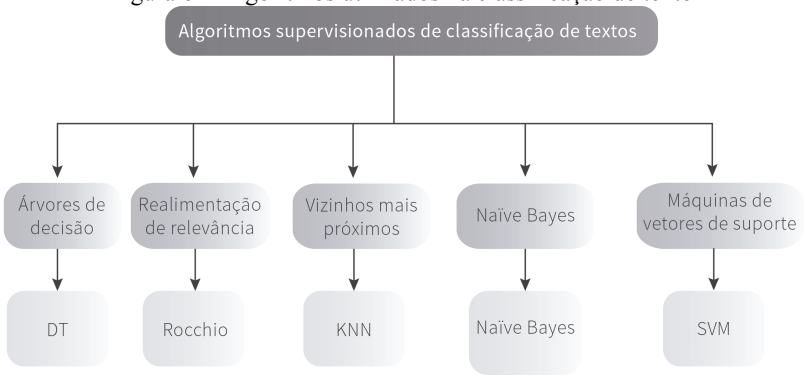
De acordo com Azam e Yao (2012), a seleção de características baseada na frequência do termo no documento permanece como uma técnica eficaz e eficiente na categorização de texto. Além disso, a identificação das características que podem reduzir a dimensão do espaço vetorial sem degradar o desempenho do classificador é amplamente utilizada na categorização texto. Na pesquisa de Yang et al. (2012) propôs-se um algoritmo de seleção de características que mede de forma abrangente o significado de um termo, tanto para a categoria como entre as categorias.

## 2.4 ALGORITMOS DE APRENDIZAGEM

Na literatura são encontrados diferentes algoritmos de aprendizagem que diferem principalmente na abordagem do processo de treinamento: paradigma supervisionado ou paradigma não supervisionado. O processo de classificação de textos normalmente é tratado no paradigma supervisionado. Ou seja, o treinamento dos dados

acontece a partir de um conjunto de entrada com seu respectivo conjunto de saída. A Figura 6 apresenta uma estrutura com os principais algoritmos de aprendizagem supervisionados, utilizados na classificação de texto, que são descritos nesta pesquisa.

Figura 6 – Algoritmos utilizados na classificação de texto




Fonte: Adaptada de Baeza-Yates e Ribeiro-Neto (2011)

### 2.4.1 Árvore de decisão (AD)

Os algoritmos de árvore de decisão (AD) (*Decision tree* – DT) são implementados para apoiar a tomada de decisão e representados por meio de árvores com um modelo de decisões possíveis e suas consequências. Esses algoritmos de AD também são utilizados como uma forma de identificar uma estratégia com maior probabilidade de atingir a meta. Existem vários algoritmos para as ADs, como por exemplo: algoritmo de Hunt, CART, ID3, C4.5, SLIQ, SPRINT; porém, o algoritmo baseado em árvore de decisão mais utilizado é o C4.5, que tem sua origem no ID3 (LOMAX; VADERA, 2013).

Segundo Baeza-Yates e Ribeiro-Neto (2011), na classificação de um novo documento, a árvore de decisão é percorrida casando os termos do documento com os termos relacionados com as arestas na árvore. Em qualquer nível de travessia da árvore, se o documento satisfaz o predicado da aresta, ele é incluído no subconjunto associado com tal aresta. A travessia recursiva é repetida até que um nodo folha seja alcançado. A classe associada com a folha torna-se a classe do documento.

Nesses algoritmos a escolha dos atributos a serem utilizados pela árvore se dá a partir de informações de entropia e ganho de informação.

Para medir o nível de informação de um atributo, utiliza-se o conceito de entropia da Teoria da Informação. O valor da entropia corresponde à impureza do atributo, e o ganho de informação é a variação da impureza.

Os métodos baseados em árvores, para classificação, dividem o espaço de entrada em regiões disjuntas para construir uma fronteira de decisão. As regiões são escolhidas com base em uma otimização heurística, em que a cada passo os algoritmos selecionam a variável que provê a melhor separação de classes de acordo com alguma função de custo. O algoritmo de classificação ID3 tem sua implementação conforme o esboço da estrutura do pseudocódigo (Figura 7).

Figura 7 – Pseudocódigo do algoritmo ID3

**Passo 1**:
**Se** todos os dados estão classificados em alguma das classes
**Então** parar ;
**Se não**, selecionar (utilizando alguma heurística) algum atributo A com valores $v_1$, $v_2$,…, $v_n$ e criar um nó de decisão.

**Passo 2**: particionar o conjunto de dados de treino T, em subconjuntos $t_1$, $t_2$, …, $t_n$ de acordo com os valores do atributo A.

**Passo 3**: aplicar o algoritmo recursivamente para cada conjunto de dados ti.

Fonte: Adaptada de Guarda (2000)

Nas pesquisas de Maazouzi e Bahi (2012) descreveu-se uma nova abordagem para mineração de dado com árvores de decisão. A proposta dos autores foi construir um modelo de AD em multicamadas, denominado MDT, nas quais cada camada é composta por várias árvores de decisão. O objetivo da abordagem MDT é melhorar o classificador de AD tradicional. O desempenho desse classificador MDT foi comparado com alguns algoritmos de AD como, por exemplo, o C4.5, e seus resultados mostram algumas melhorias com a utilização dessa abordagem.

No trabalho de Torres-Niño et al. (2013) apresentou-se um sistema para melhorar a precisão das ADs utilizando técnicas de *cluster*. Esse sistema é composto por três módulos: algoritmos de *cluster*, árvores de decisão e um módulo opcional para identificar os parâmetros adequados para o algoritmo de *cluster*. Esses três módulos trabalham em

conjunto para aumentar a precisão dessas abordagens. Realizou-se a validação dos resultados utilizando um conjunto de dados conhecidos e dois algoritmos de AD. Os percentuais de precisão do sistema proposto foram comparados com o uso exclusivo de algoritmos de *clusters* e apresentam melhorias com a utilização em conjunto dessas abordagens.

### 2.4.2 Rocchio

O algoritmo de Rocchio é um método clássico na recuperação da informação, utilizado em roteamento e filtragem de documentos. O algoritmo de Rocchio é baseado em um método de *feedback* relevante, ou realimentação de relevância, que permite a modificação de uma consulta original baseada no *feedback* do usuário. Assim, é possível produzir uma nova consulta que se aproxime melhor do interesse do usuário (MENG et al., 2013).

Segundo Baeza-Yates e Ribeiro-Neto (2011), a aplicação de Rocchio na classificação de textos é baseada na ideia de interpretar o conjunto de treinamento com informação de realimentação. Nesse caso, termos que pertencem aos documentos de treinamento de uma dada classe são ditos prover realimentação positiva, e termos que pertencem aos documentos fora da classe são ditos prover realimentação negativa. Toda informação de realimentação sobre pertencimento a classes provida no conjunto de treinamento é sumarizada em um vetor centroide no espaço de termos. Uma vez que esse centroide tenha sido computado, um novo documento de teste pode ser classificado medindo-se sua distância a esse centroide.

No trabalho de Vinot e Yvon (2003), realizou-se uma adaptação do algoritmo de Rocchio para classificação de texto utilizando a técnica de *cluster* de forma supervisionada. Nesse trabalho, o algoritmo de *cluster* é aplicado para encontrar agrupamentos dentro de classes heterogêneas. A ideia desse trabalho foi identificar subcategorias da categoria original para melhorar o processo de categorização. Os resultados dessa pesquisa não foram muito satisfatórios e variavam muito em relação à taxa de erros de acordo com a base que era utilizada no processo de classificação. Além disso, a adaptação desse método exige mais documentos na fase de treinamento do que o classificador de Rocchio tradicional. Segundo os autores, se o conjunto de exemplos fosse pequeno, a precisão do método poderia ser menor ainda.

Também na pesquisa de Gao e Guan (2012), procurou-se aperfeiçoar o algoritmo de Rocchio para a classificação de texto. Nesse trabalho foram extraídos os vetores de frequência dos termos de seus

respectivos documentos, ou seja, calculou-se para cada termo do documento seu peso tf-idf para representar o vetor do documento. Esses vetores de documentos do conjunto de treinamento geraram um modelo de classificação por meio da abordagem do algoritmo de Rocchio. Os resultados dessa pesquisa mostram que esse método que utiliza a abordagem vetorial com o algoritmo de Rocchio pode alcançar um desempenho melhor porque permite atribuir pesos aos termos tanto na realimentação positiva como na realimentação negativa.

### 2.4.3 k-Nearest Neighbors (k-NN)

O algoritmo *k-Nearest Neighbors* (k-NN), também chamado de vizinhos mais próximos, é um método não paramétrico para classificar objetos com base em exemplos de treinamento mais próximos de suas características. Esse tipo de classificador é representado por um conjunto de regras na forma normal disjuntiva que cobre melhor o conjunto de treinamento. Segundo Guo et al. (2004), para um dado documento *d* ser classificado, seus *k-Nearest Neighbors* são recuperados e constituem a vizinhança de *d*. A maioria das categorias que representam os documentos da vizinhança de *d* é recuperada e usada para decidir a classificação de *d*. Entretanto, para aplicar o k-NN é necessário definir um valor *k* apropriado, e o sucesso da classificação depende muito desse valor. Além disso, ainda segundo os autores Guo et al. (2004), o k-NN tem um custo de classificação elevado para novos casos. Isso porque quase todo o processo de computação acontece em tempo de classificação, ao invés de os exemplos de treinamento serem encontrados primeiro.

Ainda sobre pesquisas com o método de classificação k-NN, Jiang et al. (2012) argumentam que apesar de o método ser eficaz ele não é eficiente. Segundo esses autores, mesmo sendo um método simples para classificação de texto, ele apresenta três graves problemas: a complexidade computacional da similaridade na amostra é enorme; seu desempenho é facilmente afetado por amostras de treinamento com ruídos; e não constrói um modelo de classificação, uma vez que é um método de aprendizagem preguiçoso. Dessa forma, em sua pesquisa, os autores propõem um melhoramento do algoritmo k-NN para classificação de texto por meio da construção de um modelo combinado com algoritmo de *cluster*. Os resultados dessa pesquisa mostram um avanço no estado da arte do k-NN, porque em três resultados empíricos aplicados a coleções de referência o algoritmo proposto reduziu significativamente a complexidade do método.

Na pesquisa de Pang e Jiang (2013) também se apresentou uma proposta para acelerar o processo de classificação do k-NN utilizando o classificador de Rocchio. De acordo com esses autores, o classificador de Rocchio tem um desempenho eficiente, mas não consegue obter um modelo de categorização expressivo. Por isso, na proposta deles o classificador de Rocchio foi combinado com *clusters* para fortalecer a expressividade do modelo. Os resultados apresentados pelos autores mostram que eles conseguem melhorar a capacidade de categorização quando combinam essas técnicas.

### 2.4.4 Naïve Bayes

Hristea (2013) apresentou uma revisão sobre as variações no modelo de classificação e recuperação Naïve Bayes, o qual é um classificador probabilístico simples, baseado na aplicação de teorema de Bayes com suposições fortes de independência. Um termo mais descritivo para esse modelo de probabilidade seria "modelo de características independentes". Um classificador Naïve Bayes assume que a presença (ou ausência) de uma característica particular de uma classe não está relacionada com a presença (ou ausência) de qualquer outra característica. Ou seja, cada característica é analisada independentemente.

Segundo Baeza-Yates e Ribeiro-Neto (2011), classificadores probabilísticos atribuem a cada par documento-classe $[d_j, c_p]$ uma probabilidade de que o documento pertença a tal classe. Uma vez que as probabilidades tenham sido computadas para todos os pares de documento-classe que incluem o documento $d_j$, o classificador atribui a $d_j$ as classes com as maiores estimativas de probabilidade.

Assim, em um classificador Naïve Bayes baseado no modelo probabilístico clássico, um documento $d_j$ é representado por um vetor de pesos binários indicando presença ou ausência de termos de índices, como se observa:

$$\vec{d_j} = (w_{1,j}, w_{2,j}, \dots, w_{i,j}) \qquad \text{(Equação 8)}$$

Onde $w_{i,j} = 1$ se o termo $k_i$ ocorre no documento $d_j$ e $w_{i,j} = 0$, caso contrário. Para cada par documento-classe $[d_j, c_p]$, o classificador atribui um escore $S(d_j, c_p)$ dado pela relação:

$$S(d_j, c_p) = \frac{P(c_p/\vec{d}_j)}{P(\bar{c}_p/\vec{d}_j)} \qquad \text{(Equação 9)}$$

Onde $P(c_p/\vec{d}_j)$ é a probabilidade que o documento $d_j$ pertença à classe $c_p$ e $P(\bar{c}_p/\vec{d}_j)$ é a probabilidade que o documento $d_j$ não pertence à classe $c_p$. As classes com os maiores escores são atribuídas ao documento $d_j$.

Também houve propostas de melhorias no algoritmo Naïve Bayes. Segundo Balamurugan et al. (2011), o algoritmo Naïve Bayes pode falhar quando a probabilidade dos atributos, no conjunto de treinamento para cada categoria, é distribuída igualmente. Ou seja, como esse algoritmo trata a probabilidade de ocorrência dos termos independentemente, pode acontecer aleatoriamente uma distribuição de probabilidades iguais para os atributos do conjunto de treinamento. Dessa forma, o processo de classificação entre diferentes categorias ficaria prejudicado. Assim, Balamurugan et al. (2011) propõem uma melhoria no algoritmo de Naïve Bayes tradicional, denominada em sua pesquisa NB+, que calcula um fator de influência do atributo para a categoria. Consequentemente, eles tratam essa restrição que pode acontecer no Naïve Bayes tradicional e que gera probabilidades equivalentes para atributos de categorias distintas, permitindo que o algoritmo classifique corretamente nesses casos. Os resultados desse trabalho evidenciam uma melhora no estado da arte desse algoritmo.

### 2.4.5 Support Vector Machines (SVM)

O classificador *Support Vector Machines* (SVM) é um classificador linear binário não probabilístico. O SVM padrão determina, para cada conjunto de dados de entrada, de quais duas possíveis classes a entrada faz parte. De acordo com Baeza-Yates e Ribeiro-Neto (2011), o SVM constitui um método espaço-vetorial para problemas de classificação binária. Os termos de índice compõem um espaço *t*-dimensional, no qual os documentos são representados como pontos (ou vetores). Dadas as representações vetoriais para os documentos, a ideia é encontrar uma superfície de decisão que pode ser usada para melhor separar os elementos em duas classes. O hiperplano, que é aprendido a partir de dados de treinamento, divide o espaço em duas regiões. Assim, um novo documento $d_j$ pode ser classificado pela computação de sua posição relativa ao hiperplano.

Na pesquisa de Hao, Ying e Longyuan (2009) implementou-se um algoritmo de classificação Naïve Bayes combinado com *Support Vector Machines* (SVM). Segundo esses autores, o SVM tradicional apresenta uma baixa precisão em grandes amostras de treinamento. No intuito de melhorar a precisão do classificador SVM, foi proposta uma combinação do algoritmo de Naïve Bayes com SVM para reduzir o número de amostras no conjunto de treinamento. Os resultados apresentados nessa pesquisa mostram que essa combinação, além de mais confiável, melhora a precisão da classificação quando comparada com o algoritmo SVM tradicional.

## 2.5 CLASSIFICAÇÃO COM LÓGICA DIFUSA

A lógica difusa é um método para raciocínio com expressões lógicas que descrevem a pertinência a conjuntos difusos. A teoria dos conjuntos difusos pode ser entendida como um meio de especificar o quanto um objeto satisfaz a uma descrição vaga, ou seja, um conjunto difuso não tem limites precisos. Por essa razão, a teoria dos conjuntos difusos não é um método para raciocínio incerto. Em vez disso, a teoria dos conjuntos difusos trata um termo linguístico como um predicado difuso com um valor no intervalo de 0 a 1 (RUSSELL; NORVIG, 2009).

Segundo Zadrozny e Kacprzyk (2006), a lógica *fuzzy* permite representar e quantificar informações imprecisas. Ela prevê uma representação e um processamento de informações de um modo mais flexível, porque é capaz de considerar a ativação de uma variável em mais de uma função de pertinência. O uso de operadores de agregação clássicos relacionados com conectivos lógicos (AND, OR) e quantificadores "para todos", "não existe", são muitas vezes rígidos. Os quantificadores linguísticos: "quase todos", "muito mais do que 50%", são de um tipo gradual, e são mais bem modelados no contexto da lógica *fuzzy*.

Ainda de acordo com os autores Zadrozny e Kacprzyk (2006), a tarefa de categorização de textos apresenta alguma imprecisão. Além disso, pode-se esperar que os resultados da classificação sejam ambíguos. Nesse contexto, a abordagem de lógica *fuzzy* é útil porque sabe como tratar aspectos que envolvem incerteza por imprecisão na resolução de um problema.

Nesta pesquisa, o modelo de organização de documentos proposto considera que o processo de CADT envolve imprecisão porque atribui um grau de pertencimento do documento a uma ou mais

categorias. Isso normalmente se justifica porque um documento pode não pertencer unicamente a uma categoria. Ou seja, existe a necessidade de implementar classificadores capazes de lidar com a imprecisão e a subjetividade do processo de CADT. Outros trabalhos relacionados a esta pesquisa também tratam da possibilidade de aplicar elementos da lógica *fuzzy* para fins de recuperação da informação (BORDOGNA; PASI, 2005; HERRERA-VIEDMA, 2001; KACPRZYK et al., 2000; KARABULUT, 2013; SHAREF; KASMIRAN, 2012; ZADROZNY; KACPRZYK, 2006).

As pesquisas iniciais, que aplicavam a teoria de conjuntos *fuzzy* para a recuperação da informação, normalmente utilizavam consultas booleanas que permitiam atribuir pesos aos termos, tanto na indexação quanto na representação da consulta. Assim, cada peso poderia ser descrito como uma função de pertinência, indicando o quanto o documento em questão era parte da coleção total.

Apesar do esforço em introduzir pesos numéricos para melhorar ambas as representações de documentos e a linguagem de consulta, a utilização de pesos requeria um conhecimento claro da semântica da consulta, a fim de traduzir um conceito difuso em um valor numérico preciso. Além disso, era difícil modelar a correspondência entre as consultas e os documentos de forma a preservar a semântica das consultas dos usuários.

Segundo Kacprzyk et al. (2000), o processo de indexação adotado pelos sistemas de recuperação da informação produz representações formais dos documentos, enquanto que as consultas são especificadas diretamente pelos usuários por meio de linguagem de consulta no sistema. O mecanismo que realiza a correspondência entre essas duas representações com o objetivo de estimar documentos relevantes para a consulta não é simples. É um mecanismo complexo e permeado por imprecisão e incerteza.

Ainda segundo Kacprzyk et al. (2000), uma direção promissora para melhorar esses sistemas é um modelo de subjetividade intrínseca na interpretação e seleção de informações, ou seja, um modelo capaz de aprender conceito de relevância dos usuários. A teoria dos conjuntos *fuzzy* proporciona um suporte adequado e natural para definir os sistemas de recuperação de informação tolerantes à imprecisão. Na pesquisa de Kacprzyk et al. (2000), empregou-se a teoria de conjuntos difusos para generalizar consultas de linguagem booleanas, definindo uma linguagem de consulta flexível. Assim, descritores linguísticos são formalizados por meio de variáveis linguísticas e novos operadores de

agregação (*at least, most of*) são definidos com seus significados próprios e entre os conectivos AND e OR.

Herrera-Viedma (2001) propôs um modelo de recuperação de informação (RI) usando uma abordagem linguística *fuzzy* ordinal. Esse modelo aceita consultas linguísticas ordinais em função de dois elementos de ponderação: os termos da consulta e as subexpressões. Dessa forma, segundo o autor, os usuários podem facilmente expressar simultaneamente várias restrições semânticas em uma consulta. Ainda de acordo com o autor, o uso desse modelo não é simples, e os usuários não especialistas podem se confundir quando desejam aplicar diferentes associações semânticas em um mesmo elemento da consulta ao mesmo tempo.

Bordogna e Pasi (2005) consideram o problema da indexação de documentos heterogêneos estruturados e a recuperação de documentos semiestruturados. Esses autores propõem um paradigma flexível tanto para indexação de documentos como para formulação de consultas. Ao nível de indexação é proposto um modelo que constrói representações dos documentos a partir do ponto de vista dos usuários. Ainda segundo esses autores, uma abordagem baseada nessa indexação personalizada constitui uma maneira de projetar sistemas de RI flexíveis que são capazes de aprender por meio das preferências dos usuários. Ou seja, eles permitem que os usuários especifiquem suas preferências em seções do documento que eles estimam ter mais informações de interesse, bem como quantificam linguisticamente o número das seções que determinam o potencial interesse global pelos documentos. Para a formulação das consultas é proposta uma linguagem flexível para expressar condições de seleção suaves sobre a estrutura dos documentos. Por exemplo: o quantificador linguístico "mais" especifica uma restrição suave sobre a estrutura dos documentos, não permitindo que documentos de potencial interesse sejam ignorados. Dessa forma, essa linguagem de consulta flexível difusa permite selecionar um subconjunto de documentos semiestruturados a partir de uma coleção heterogênea.

Na pesquisa de Zadrozny e Kacprzyk (2006), propôs-se uma extensão do modelo booleano, que pode ser considerada um ponto de partida para utilizar a lógica *fuzzy* em RI. Em seu trabalho abordaram-se questões relacionadas com a representação e a classificação de documentos. Assim, aplicaram-se cálculos clássicos de proposições com quantificadores linguísticos propostos por Zadeh, diretamente nas consultas. Segundo esse autor, o conceito de quantificador linguístico ("maior que", "quase todos", "muitos") pode ser diretamente empregado

na definição de um conjunto *fuzzy*, tornando o quantificador mais ajustável. Além disso, Zadrozny e Kacprzyk (2006) empregaram resultados relacionados à consulta difusa em RI apresentando várias interpretações para os seus termos.

A maior parte dos trabalhos iniciais que trataram da inserção das técnicas de lógica *fuzzy* se concentra em pesquisas relacionadas com RI. Mas existem também trabalhos que buscam adaptar seus processos de classificação de documentos por meio da lógica *fuzzy*.

Assim, no trabalho de Sharef e Kasmiran (2012) apresentou-se uma gramática difusa como uma técnica para construir um classificador de texto. Essa gramática difusa foi comparada com outros métodos de aprendizado de máquina, tais como tabela de decisão, *support vector machines*, métodos estatísticos e k-NN. Os resultados apresentados em seu trabalho mostraram que essa gramática *fuzzy* obteve resultados promissores em relação a outros métodos de aprendizagem de máquina de referência.

Na pesquisa de Karabulut (2013) consideraram-se duas fases principais para a categorização: a redução de termos e a classificação. Dessa forma, os autores propõem uma estratégia para redução de termos na categorização de documentos. Para avaliar o desempenho dessa abordagem com redução de termos, foi proposto um novo classificador denominado FURIA. Esse classificador foi implementado com algoritmos de regra de indução e lógica *fuzzy*. Seus resultados foram comparados com os algoritmos de Naïve Bayes e SVM e apresentaram uma eficácia na tarefa de ordenação dos documentos classificados.

## 2.6 CLASSIFICAÇÃO EM MÚLTIPLAS CATEGORIAS

A classificação tradicional trabalha com a aprendizagem de um conjunto de exemplos que estão associados a uma única categoria ou classe e trata o processo de aprendizagem como um problema de classificação binária. Na pesquisa elaborada por Tsoumakas e Katakis (2007) apresentou-se uma visão geral sobre métodos de classificação em múltiplas categorias com comentários sobre os pontos fortes e fracos de cada método. Segundo esses autores, a literatura sobre os métodos de classificação múltipla ainda é escassa, porque a maior parte dos algoritmos de classificação se concentra em uma única categoria.

Os métodos para a classificação em múltiplas categorias são reunidos em dois grupos principais: os métodos de transformação de problema e os métodos de adaptação de algoritmo. No primeiro grupo, o

problema é decomposto em vários problemas binários que farão a classificação dos documentos independentemente. O resultado final é determinado pela agregação dos resultados de todos os classificadores binários independentes (CHERMAN; MONARD; METZ, 2011).

Os métodos do segundo grupo estendem os algoritmos de aprendizagem para lidar com múltiplas categorias diretamente (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010).

### 2.6.1 Métodos de transformação de problema

De acordo com Tsoumakas, Katakis e Vlahavas (2010), o método de transformação de problema mais comum é o denominado PT4, também conhecido como relevância binária (*binary relevance* – BR). Esse classificador binário rotula o conjunto de dados originais como pertencendo ou não à classe considerada. Ou seja, para cada exemplo existe sua classificação da seguinte forma: $X \rightarrow \{l, \neg l\}$, para cada categoria avaliada. Esta é uma solução que lida com o problema de múltiplas categorias com um classificador binário.

Para exemplificar o uso dos métodos de transformação de problema, considere os exemplos de quatro documentos classificados em mais de uma categoria (Quadro 1).

Quadro 1 – Conjunto de textos em múltiplas categorias

| Texto | Esporte | Ciência | Política | Educação |
|---|---|---|---|---|
| 1 | X | | | X |
| 2 | | | X | X |
| 3 | X | X | | |
| 4 | | X | X | |

Fonte: Elaborado pela autora

No Quadro 2 apresentam-se os quatro conjuntos obtidos por meio do uso do método de transformação de problema PT4 no conjunto de exemplos apresentados no Quadro 1.

Quadro 2 – Resultados obtidos pelo uso do PT4

| Texto | Esporte | ¬Esporte |
|---|---|---|
| 1 | X | |
| 2 | | X |
| 3 | X | |
| 4 | | X |

| Texto | Ciência | ¬Ciência |
|---|---|---|
| 1 | | X |
| 2 | | X |
| 3 | X | |
| 4 | X | |

| Texto | Política | ¬Política |
|---|---|---|
| 1 | | X |
| 2 | X | |
| 3 | | X |
| 4 | X | |

| Texto | Educação | ¬Educação |
|---|---|---|
| 1 | X | |
| 2 | X | |
| 3 | | X |
| 4 | | X |

Fonte: Elaborado pela autora

Segundo Tsoumakas, Katakis e Vlahavas (2010), outro método de transformação de problema é o *Label Powerset* (LP). Nesse método, cada subconjunto de diferentes categorias envolvidas na classificação de um documento é considerado como uma única categoria para a tarefa de classificação. Por exemplo, se uma instância é associada com três categorias C1, C2, C4, então a nova categoria será $C_{1,2,4}$. Dessa forma, todo o conjunto de categorias que ocorre em um documento (instância) é transformado para uma única categoria que considera as demais. Assim, o conjunto fica pronto para o processo de treinamento e classificação em um classificador *single-label*. O Quadro 3 apresenta o conjunto de dados transformado a partir do método LP em relação ao conjunto de exemplos do Quadro 1.

Quadro 3 – Conjunto de categorias transformado usando o método LP

| Texto | Categoria |
|---|---|
| 1 | $\lambda_{esporte,educação}$ |
| 2 | $\lambda_{política,educação}$ |
| 3 | $\lambda_{esporte,ciência}$ |
| 4 | $\lambda_{ciência,política}$ |

Fonte: Elaborado pela autora

Segundo Read (2008), outro método de transformação de problema que estende o *Label Powerset* (LP) é o método de transformação de problema com poda (do inglês, *Pruned Problem Transformation* – PPT), também referenciado como *Pruned Set* (PS). O método PS remove as limitações do LP podando os conjuntos de categorias que estão ocorrendo menos do que um limite definido pelo usuário. Assim, uma categoria que aparece poucas vezes no conjunto de treinamento é desconsiderada.

Na pesquisa de Modi e Panchal (2012) avaliou-se o desempenho dos métodos de transformação de problemas considerando características como precisão e correspondência exata. Em sua abordagem foram considerados os métodos: relevância binária (BR), Label Powerset (LP), e o método *Pruned Set* (PS). Seus resultados, obtidos em conjuntos reais, mostraram que os métodos LP e PS podem dar melhores resultados do que o método BR.

Uma das vantagens do método LP em relação ao BR é que ele considera as correlações entre as categorias. Porém, se houver um grande conjunto de categorias, esse número pode crescer exponencialmente depois da aplicação do método LP. Assim, pode-se apresentar categorias com poucos exemplos associados no conjunto de treinamento. Ainda de acordo com Modi e Panchal (2012), isso pode resultar em um processo de aprendizagem difícil, fornecendo categorias em desequilíbrio na fase de treinamento. Além disso, o LP tem uma limitação quanto à capacidade de previsão em uma determinada categoria, porque só pode realizar isso confiavelmente quando possui essa categoria em seu conjunto de treinamento. Esse método também é referenciado na literatura como *Label Combination* (LC).

Existem na literatura outros métodos de transformação de problema, como por exemplo: *Random k-labelsets* (RAkEL) e *Ranking by Pairwise Comparision* (RPC); porém, esses métodos não são descritos neste trabalho porque se afastam do escopo desta proposta.

### 2.6.2 Métodos de adaptação de algoritmo

Algumas pesquisas tratam do desenvolvimento de adaptação de algoritmos de aprendizagem para classificação em múltiplas categorias. Entre elas, incluem-se propostas de adaptação do *k-Nearest Neighbors* (k-NN) e do *Support Vector Machines* (SVM).

O ML-kNN (ZHANG; ZHOU, 2007) é uma adaptação do algoritmo de aprendizagem k-NN para múltiplas categorias. Na verdade, ele segue o paradigma do método de transformação de problema BR. Em essência, a ML-kNN utiliza o algoritmo k-NN independentemente para cada categoria. Ele encontra os exemplos mais próximos de k para a instância de teste e os considera da categoria, considerando o restante como se não fossem. O que diferencia essencialmente esse método a partir da aplicação do algoritmo k-NN original para o problema transformado utilizando o BR é o uso do princípio máximo a posteriori (*maximum a posteriori* – MAP). Depois que são identificados os k vizinhos mais próximos no conjunto de treinamento, para cada documento desconhecido é aplicado o princípio do MAP. Esse princípio determina, com base em informações estatísticas obtidas por meio do número de documentos vizinhos pertencentes a cada categoria possível, a qual conjunto de categorias a instância do documento desconhecido pertence. Assim, ML-kNN tem também a capacidade de produzir um ranqueamento das categorias em sua saída.

Luo e Zincir-Heywood (2005) apresentam dois sistemas de classificação de documentos em múltiplas categorias, que também são baseados no classificador k-NN. A principal contribuição de seu trabalho está na fase de pré-processamento para a representação efetiva dos documentos. Para a classificação de uma nova instância, os sistemas inicialmente encontram os exemplos mais próximos de k. Em seguida, para cada ocorrência de categoria encontrada no conjunto de exemplos mais próximos de k é incrementado um contador correspondente a essa categoria. Por fim, as N categorias com as maiores contagens são apresentadas. Segundo esses autores, esta é uma estratégia inadequada para uso no mundo real, onde o número de categorias de uma nova instância é desconhecido.

Godbole e Sarawagi (2004) apresentaram duas melhorias para o classificador SVM em conjunto com o método BR para a classificação em múltiplas categorias. A ideia principal desse trabalho é ampliar o conjunto de dados originais com características extras contendo previsões para cada classificador binário. Então, em uma segunda rodada de treinamento, novos classificadores binários usariam o

conjunto de dados estendidos. Para a classificação de um novo exemplo, os classificadores binários da primeira rodada são inicialmente utilizados, e sua saída é anexada às características do exemplo para formar um metaexemplo. Esse metaexemplo é, então, classificado pelos classificadores binários da segunda rodada. Por meio dessa extensão, a abordagem leva em consideração as possíveis dependências entre as diferentes categorias. A segunda melhoria, de acordo com Godbole e Sarawagi (2004), é específica ao SVM e diz respeito aos limites deste em problemas de classificação *multi-labels*. Os limites são melhorados, removendo instâncias de treinamento negativas que estavam muito próximas e dentro da distância *threshold* para a aprendizagem. Também foram removidas instâncias de treinamento negativas de uma classe inteira, se elas eram muito similares a uma classe positiva.

Na pesquisa de Moschitti, Ju e Johansson (2012), modelaram-se dependências para categorização hierárquica de texto em múltiplas categorias. Em sua pesquisa considera-se tanto a estrutura em hierarquia como a probabilidade dos nodos individualmente. Além disso, para melhor descrever o papel da categoria nas relações, foram considerados dois casos: (i) esquemas tradicionais, em que nós-pais incluem todos os documentos de suas categorias-filhas; e (ii) os esquemas mais gerais, em que os filhos podem incluir documentos não pertencentes aos seus pais. Ainda segundo esses autores, falhas no uso dessas abordagens de categorização hierárquica em múltiplas categorias são causadas pela complexidade de modelar todas as dependências possíveis entre as categorias.

### 2.6.3 Considerações sobre os métodos

Independentemente da abordagem, os diferentes métodos para classificação em múltiplas categorias buscam distinguir a heurística utilizada para a tarefa de classificação. O método de transformação de problemas define uma classificação focada na divisão do problema de múltiplas categorias em categorias únicas. A abordagem do método de adaptação de algoritmos trata o problema de classificação em múltiplas categorias por meio da adequação dos diferentes algoritmos de aprendizagem da literatura, por exemplo: SVM, k-NN, AD, RN.

O Quadro 4 apresenta uma síntese dos diferentes trabalhos que se apresentam considerando as duas abordagens: transformação de problema e adaptação de algoritmo.

Quadro 4 – Síntese das abordagens de trabalhos em múltiplas categorias

| Método | Trabalho | Autores | Abordagem |
|---|---|---|---|
| Transformação de problema | *Multi-label classification: an overview*. | Tsoumakas, Katakis (2007) | Revisão inicial dos métodos de transformação de problemas: PT4 – BR. |
| Transformação de problema | *Mining multi-label data*. | Tsoumakas, Katakis e Vlahavas (2010) | Formalização dos métodos de transformação de problemas. |
| Transformação de problema | *Experimental comparison of different problem transformation methods for multi-label classification using MEKA*. | Modi e Panchal (2012) | Avaliação do desempenho dos métodos: BR, PS e LC. |
| Adaptação de algoritmo | *Discriminative methods for multi-labeled classification*. | Godbole e Sarawagi (2004) | SVM com PT4. |
| Adaptação de algoritmo | *Evaluation of two systems on multi-class multi-label document classification*. | Luo e Zincir-Heywood (2005) | k-NN, melhoramento do pré-processamento. |
| Adaptação de algoritmo | *A lazy learning approach to multi-label learning*. | Zhang e Zhou (2007) | k-NN com PT4. |
| Adaptação de algoritmo | *Modeling topic dependencies in hierarchical text categorization*. | Moschitti; Ju e Johansson (2012) | Categorização de texto hierárquico em múltiplas categorias. |

Fonte: Elaborado pela autora

Dos diferentes métodos descritos para a classificação em múltiplas categorias, a abordagem que mais se aproxima do método proposto por este trabalho é o método de transformação de problema de relevância binária (BR). Da mesma forma, o método implementado

neste trabalho divide o problema de classificação em múltiplas categorias em problemas de classificação em uma única categoria. Assim, esse método classifica os documentos considerando um conjunto de regras que tratam essa tarefa de forma independente.

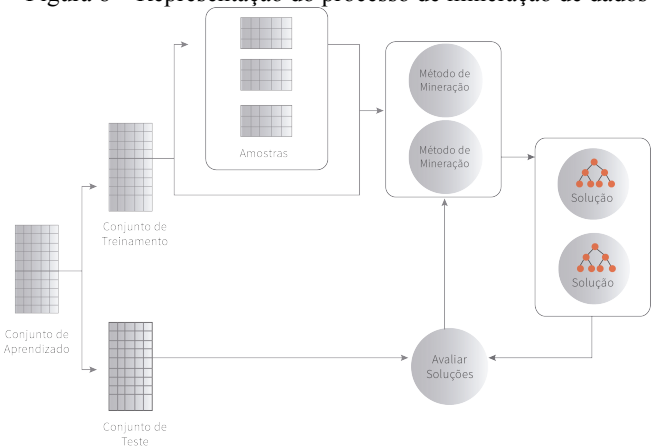
## 2.7 PROCESSO DE DESCOBERTA DE CONHECIMENTO

A mineração de dados (*data mining* – DM) está inserida em um processo maior denominado descoberta de conhecimento em banco de dados (*Knowledge discovery in database* – KDD). Mais especificamente a mineração de dados se restringe à obtenção de modelos, sendo a DM uma instância do processo de KDD (BRAGA, 2005).

Uma das principais etapas do processo de descoberta de conhecimento (KDD) é a DM. Ou seja, a mineração de dados é um processo computacional para descobrir padrões em grandes conjuntos de dados e normalmente envolve métodos de inteligência artificial, como aprendizagem de máquina e redes neurais, estatística e gerenciamento de banco de dados (CLIFTON, 2010; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Normalmente um processo de mineração de dados envolve a divisão do conjunto de aprendizado em duas partes: conjunto de treinamento e conjunto de teste. Do conjunto de treinamento são extraídas amostras às quais se aplicam métodos de mineração. O resultado da aplicação desses métodos tende a realizar o reconhecimento de padrões sobre os dados gerando um modelo. Realiza-se a avaliação do modelo obtido por meio de um conjunto de testes (Figura 8).

Figura 8 – Representação do processo de mineração de dados




Fonte: <http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-KDD-DM.pdf>

Segundo Clifton (2010), o processo completo de mineração de dados envolve muitas etapas, mas as três principais etapas computacionais são: o processo de aprendizagem do modelo, a avaliação do modelo e o uso do modelo. Ainda segundo Clifton (2010), existem muitos tipos de mineração de dados, normalmente divididos de acordo com a informação conhecida (atributos) e com o conhecimento buscado por esse modelo.

Braga (2005) descreve que as diferenças entre os tipos de mineração não são de essência, mas de apresentação e implementação. Esses sistemas passam pelas mesmas etapas: coleta de dados, depuração e análise, resultando em um "modelo descritivo", e, caso se deseje, os resultados serão utilizados na construção de um "modelo preditivo".

A modelagem descritiva divide os dados em grupos (*cluster*); no entanto, as aglomerações não são conhecidas antecipadamente, ou seja, os padrões descobertos por meio das análises dos dados são utilizados para determinar os grupos.

A modelagem preditiva é utilizada quando o objetivo é estimar o valor de um atributo particular e existem dados de treinamento para os quais são conhecidos os valores do atributo. Um exemplo de modelagem preditiva é o processo de classificação. Ainda quanto aos algoritmos utilizados no processo de classificação da modelagem preditiva, de acordo com Dziekaniak (2010), existem três abordagens mais conhecidas: a abordagem simbólica, baseada em árvores de decisão; a abordagem biológica, implementada por redes neurais e algoritmos

genéticos; e a abordagem estatística, realizada por algoritmos de *Naïve-Bayes* e *k-Nearest Neighbors*.

Essas abordagens buscam analisar um conjunto de dados classificados (em classes) e desenvolver uma descrição ou modelo para cada classe utilizando os atributos dos dados. Assim, por meio do modelo descoberto é possível prever a classe (o valor do atributo meta) de novos dados.

Por outro lado, a classificação também procura descobrir um relacionamento entre os atributos previsores e o atributo meta. Para isso, são utilizados registros cujas classes são conhecidas, para que na construção de um modelo possam ser identificados os objetos não classificados. Assim, é possível classificá-los e estabelecer uma previsão a partir do modelo (PAPPA; FREITAS, 2009).

Além disso, de acordo com Barth (2009), o uso de algoritmos de aprendizagem de máquina permite a criação de estruturas simbólicas que são compreensíveis por pessoas. Assim, é possível entender quais atributos podem ser mais significativos na ativação das regras e como utilizá-los da melhor forma.

De fato, o uso de algoritmos de aprendizado de máquina com abordagem simbólica traz inúmeros benefícios para a construção da modelagem difusa e suas regras. Sabe-se que existem diversos algoritmos de aprendizagem de máquina que geram representações simbólicas. Segundo Russell e Norvig (2009), alguns são mais expressivos que outros, como, por exemplo, a lógica de primeira ordem é mais expressiva do que a lógica proposicional. Porém, quanto mais expressiva a linguagem adotada para a representação de uma hipótese, maior é o custo computacional que os algoritmos de aprendizagem de máquina têm para gerar essas hipóteses.

Segundo Hastie, Tibshirani e Friedman (2009), aplicações de mineração de dados industriais e comerciais tendem a ser especialmente difíceis em termos das exigências dos procedimentos de aprendizagem. Os conjuntos de dados são geralmente muito grandes em relação ao número de entradas e ao número de variáveis para cada entrada. Assim, considerações computacionais têm um papel importante na escolha do método de aprendizagem. Além disso, deve-se considerar que os dados normalmente estão embaralhados, ou seja, as amostras apresentam dados misturados: quantitativo, binário e variáveis categóricas, tendo estas últimas, muitas vezes, vários níveis. Em geral, as amostras apresentam falta de valores, e observações completas são raras. No Quadro 5 apresentam-se sínteses sobre as características de alguns

métodos de aprendizagem (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Quadro 5 – Características de alguns métodos de aprendizagem

| Característica dos métodos | Rede neural | SVM | Árvores de decisão | k-NN |
|---|---|---|---|---|
| Manipulação natural com tipos de dados misturados | Fraco | Fraco | Bom | Fraco |
| Ajuste com valores faltantes | Fraco | Fraco | Bom | Bom |
| Robustez a *outliers* no espaço de entrada | Fraco | Fraco | Bom | Bom |
| Escalabilidade computacional (N grande) | Fraco | Fraco | Bom | Fraco |
| Capacidade de lidar com entradas irrelevantes | Fraco | Fraco | Bom | Fraco |
| Capacidade de extrair combinações de características lineares | Bom | Bom | Fraco | Razoável |
| Interpretabilidade | Fraco | Fraco | Razoável | Fraco |

Fonte: Hastie, Tibshirani e Friedman (2009)

De acordo com o Quadro 5, Hastie, Tibshirani e Friedman (2009) avaliam que as árvores de decisão possuem como base as melhores características para impulsionar a aprendizagem de dados em aplicações de mineração. De acordo com Barth (2009), as árvores de decisão utilizam sentenças proposicionais que conseguem representar uma decisão complexa a partir de diversas decisões elementares. Segundo ele, o custo computacional que um algoritmo indutor dessas árvores tem para gerá-las é baixo.

## 2.8 CONSIDERAÇÕES SOBRE O ESTADO DA ARTE

O processo de revisão da literatura buscou identificar o estado da arte das pesquisas relacionadas com o modelo de organização de documentos proposto. Como o modelo proposto é embasado em um método de classificação de documentos com *fuzzy* em múltiplas

categorias, foram analisados artigos relacionados com o processo de categorização de documentos texto, considerando abordagens de recuperação da informação, algoritmos de aprendizagem, classificação com lógica *fuzzy* e classificação em múltiplas categorias.

A revisão da literatura considerou artigos publicados em conferências e em *journals* indexados em bibliotecas digitais da área. As bases de dados utilizadas nesta pesquisa foram: IEEE Xplore, ACM Digital Library, Springer Link e ScienceDirect.

Nesta pesquisa buscou-se identificar o progresso do estado da arte dos procedimentos e algoritmos que tratam sobre categorização de textos entre 2000 e 2013. Assim, os termos de busca se concentraram na combinação de *strings* e operadores lógicos formando a seguinte expressão:

*(“text categorization” OR “text classification” OR "text document categorization ") AND (“retrieval information” OR “learning algorithms” OR “fuzzy" OR “multi-label”)*

Os procedimentos e critérios para seleção dos artigos mais relevantes a esta pesquisa consideraram aspectos como a natureza da proposta, o método desenvolvido e as tecnologias semelhantes relacionadas com esse modelo. A fase inicial da revisão da literatura considerou um total de 627 artigos distribuídos nas quatro bases de referências citadas (Tabela 1).

Tabela 1 – Total de artigos avaliados por base

| Base | Total |
|---|---|
| ACM | 106 |
| Springer Link | 203 |
| IEEE Xplore | 249 |
| ScienceDirect | 69 |
| Total de artigos | 627 |

Fonte: Elaborada pela autora

Após uma primeira leitura dos títulos, resumos e palavras-chave, consideraram-se 112 artigos. Em uma leitura detalhada dos artigos, levaram-se em conta 21 artigos mais próximos com a proposta desta tese. Todos esses artigos foram referenciados ao longo deste capítulo.

Das diferentes citações ao longo deste capítulo, estruturaram-se as pesquisas que mais contribuíram e que tinham propostas parecidas com as deste trabalho. No Quadro 6 é possível identificar o título, os autores, os destaques do trabalho e as semelhanças com esta pesquisa.

Quadro 6 – Trabalhos relacionados

| Título | Autores | Destaques | Semelhanças com esta pesquisa |
|---|---|---|---|
| Text document categorization by term association | Antonie e Zaiane (2002) | Aplicação de mineração de regras de associação para implementação de um classificador associativo. | Construção da etapa de pré-processamento dos documentos.<br><br>Uso de mineração, porém no contexto de regras associativas. |
| Computing with words for text processing: an approach to the text categorization | Zadrożny e Kacprzyk (2006) | Aplicação de uma extensão do modelo booleano (RI). | Uso da lógica *fuzzy*. Considera um grau de pertencimento à categoria na classificação. |
| Mining multi-label data | Tsoumakas, Katakis, Vlahavas (2010) | Definição do método de transformação de problema – PT5. | Distribuição de graus de certeza para todas as categorias |
| Comparison of term frequency and document frequency based feature selection metrics in text categorization | Azam e Yao (2012) | Algoritmos de seleção de características. | Uso de métricas baseadas na frequência do termo, interessante para conjuntos de características menores. |
| Examining text categorization methods for incidents analysis | Sharef e Kasmiran (2012) | Uso de uma gramática difusa para construção do classificador. | Uso da lógica *fuzzy* para implementação do classificador. |
| Exploring Feature Selection and Support Vector Machine in Text Categorization | Abdul-Rahman et al. (2013) | Uso de SVM. | Etapa de pré-processamento reduzindo a dimensionalidade de documentos de textos. |
| *Fuzzy* unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection | Karabulut (2013) | Uso de estratégia para redução de termos para categorização. | Uso da lógica *fuzzy* para implementação do classificador. |
| Improving Accuracy of Decision Trees Using Clustering Techniques | Torres-Niño (2013) | Melhoramento dos resultados da AD com uso de *cluster*. | Combinação do uso de *cluster* para melhorar o desempenho das ADs geradas. |

Fonte: Elaborado pela autora

A maioria dos algoritmos de aprendizado de máquina foi projetada para a classificação em um único rótulo, isto é, um documento só pode pertencer a uma categoria (JIANG; TSAI; LEE, 2012). Na classificação de texto em múltiplos rótulos, um documento pode pertencer a mais de uma categoria. Além disso, muitos autores têm trabalhado com técnicas de lógica *fuzzy* na recuperação e classificação de documentos (BORDOGNA; PASI, 2005; HERRERA-VIEDMA, 2001; KACPRZYK et al., 2000; KARABULUT, 2013; SHAREF; KASMIRAN, 2012; ZADROZNY; KACPRZYK, 2006). Tratar o processo de CADT considerando o conceito de imprecisão é natural, porque não existem limites precisos sobre os conjuntos de categorias analisadas; além disso, os resultados de um processo de categorização podem ser ambíguos.

Um aspecto atraente na implementação das árvores de decisão, em relação aos métodos Naïve Bayes, Rocchio, *k-Nearest Neighbors e* SVM é que elas são facilmente interpretáveis e modificáveis. Apesar da complexidade do método SVM, de acordo com Lee e Kageura (2007), esse é um dos melhores métodos para a categorização de textos.

Todo processo de categorização textual é importante, e há mais demanda por modelos de classificação eficazes e eficientes que considerem, sobretudo, fatores de imprecisão. Ainda sobre o desempenho dos algoritmos de classificação, na pesquisa de Dan, Lihua e Zhaoxin (2013), observou-se em seus resultados que o desempenho não está associado somente à escolha do algoritmo, mas também a diferenças entre as bases de dados das categorias. Daí a importância de executar o pré-processamento considerando uma seleção de características que sejam relevantes para o processo de classificação.

Apesar de vários esforços no sentindo de elaborar e implementar modelos que classifiquem um grande conjunto de informação de modo eficiente, de acordo com a literatura ainda existem lacunas na integração de propriedades que são significativas nesse processo. Segundo Antonie e Zaiane (2002) e Kim et al. (2009), algumas dessas propriedades se referem a procedimentos que são mais precisos, outros que oferecem um modelo de classificação mais interpretável. Além disso, algumas pesquisas ainda focalizam somente a classificação em um único rótulo e outras nem consideram características relacionadas à imprecisão no problema de CADT. De acordo com Zadrozny e Kacprzyk (2006), a abordagem de lógica *fuzzy* provou ser útil nesse tipo de contexto.

Ming, Lu e Chuan-Bo (2011) afirmam que o objetivo da classificação de texto é rotular documentos em classes temáticas por meio de um conjunto predefinido. Ainda de acordo com esses autores,

diversos métodos têm sido aplicados para a tarefa de classificação de texto, incluindo a abordagem do *k-Nearest Neighbors* (k-NN) (BANG; YANG J. D.; YANG H. J., 2006; TAN, 2006), as abordagens *Naïve Bayes* (HRISTEA, 2013), *Support Vector Machines* (HAO; YING; LONGYUAN, 2009) e Árvores de Decisão (LOMAX; VADERA, 2013).

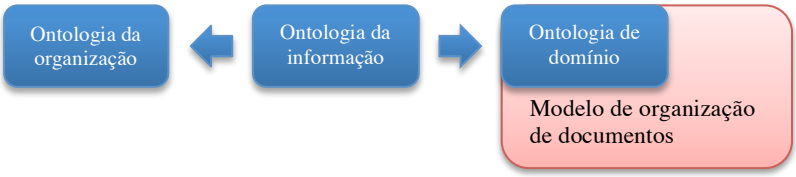
# 3 MODELO PARA ORGANIZAÇÃO DE DOCUMENTOS

O modelo proposto permitirá que uma coleção de documentos da MO seja classificada em função de categorias de interesse da organização. Para isso, aplica-se um método de classificação de documentos considerando múltiplas categorias. O foco desse modelo é atingir melhorias nos processos de gestão dos artefatos de conhecimento organizacional, permitindo que documentos específicos sejam acessados pelas pessoas interessadas e que documentos de interesse comum sejam percebidos pelos diferentes indivíduos da organização.

## 3.1 ESCOPO DO MODELO

A especificação funcional desse modelo de organização de documentos está inserida no contexto da ontologia de domínio (Figura 9). Assim, esse modelo interage por meio dessa ontologia, pois ela detém a descrição do conteúdo das fontes de informação utilizadas no contexto da organização, considerando as informações, palavras-chave e frases de um documento. Esse modelo trata a organização de documentos no escopo da língua portuguesa.

Figura 9 – Escopo do modelo de organização de documentos

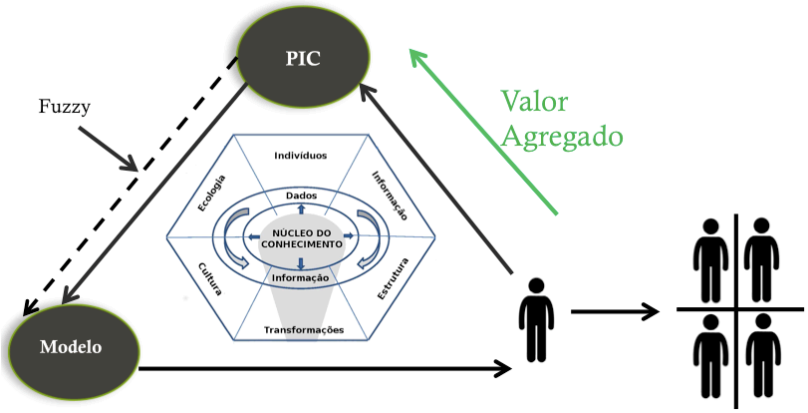



Fonte: Elaborada pela autora

O modelo de organização desenvolvido é parte de uma pirâmide que considera os processos intensivos em conhecimento (PIC) atuando diretamente na construção desse modelo. A inferência *fuzzy* possibilita que o conhecimento da MO, no centro da pirâmide, seja potencializado e identificado a partir dos interesses do indivíduo. Dessa forma, o indivíduo pode interagir por meio dos PICs com o objetivo de identificar novas categorias de interesse (Figura 10). Além disso, o trabalho do indivíduo migra para um nível coletivo e não mais individual, porque se

pode tratar as informações de interesse que são comuns aos grupos dentro da organização.

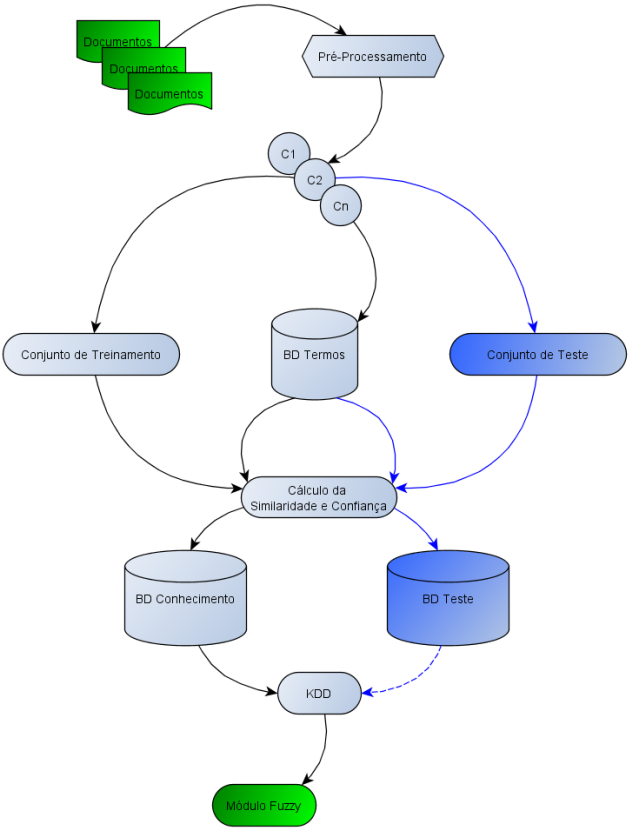
Figura 10 – Pirâmide de interação PIC, modelo e indivíduo




Fonte: Elaborada pela autora

Para efetivar a implementação desse modelo de organização, foi desenvolvido um método de classificação de documentos. Esse método considera a alta dimensionalidade envolvida na classificação dos documentos, ou seja, como um documento contém muitas informações que podem ser representadas em várias classes, ele é categorizado em múltiplas categorias. A Figura 11 apresenta os passos do desenvolvimento do modelo de organização de documentos com base nesse método.

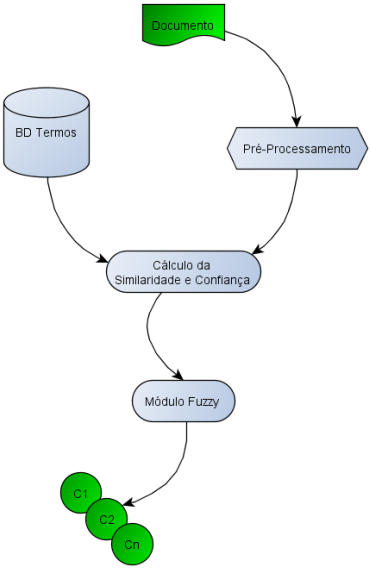
Figura 11 – Método para o desenvolvimento do modelo




Fonte: Elaborada pela autora

O método proposto considera um processo de descoberta de conhecimento (KDD) sobre as regras que são implementadas em uma modelagem *fuzzy*. O uso da lógica *fuzzy* é capaz de tratar a sobreposição de categorias em um documento lidando com a imprecisão das variáveis de entrada. A estrutura do modelo de organização de documentos é apresentada na Figura 12.

Figura 12 – O modelo de organização de documentos




Fonte: Elaborada pela autora

A aplicação do modelo é realizada a partir de um documento pré-processado, cuja similaridade e confiança são calculadas de acordo com a definição de uma base de termos. Essas duas variáveis são as entradas da modelagem *fuzzy*. Assim, uma coleção de documentos de uma organização pode ser categorizada atribuindo sua pertinência a cada categoria identificada como de interesse para a organização.

Nas próximas seções detalha-se o processo de desenvolvimento desse modelo considerando suas etapas: definição da base de termos, definição das variáveis, definição da base de conhecimento e a implementação da modelagem *fuzzy*.

### 3.1.1 Definição da base de termos

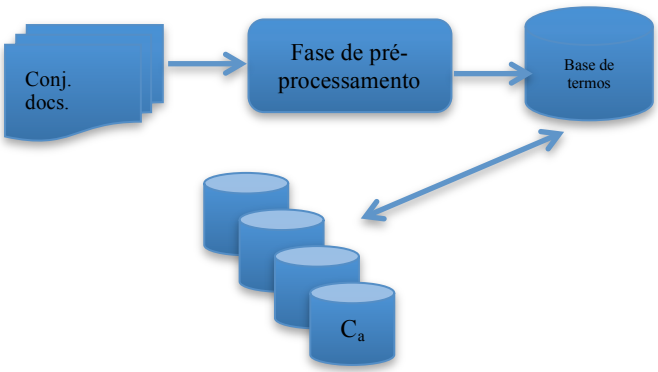
A base de termos é definida em função de um conjunto de palavras e suas respectivas categorias. Os termos são extraídos de uma coleção de documentos de uma determinada categoria considerada relevante para a organização. Dessa forma, todas as categorias de

interesse para a organização devem ser identificadas no processo de construção da base de termos.

Todos os documentos são pré-processados para a transformação em vetores de termos por categoria. Na Figura 11 a base de termos é expressa pelos "BD Termos". Essa base é construída para especificar um grau de relevância dos termos associado às categorias.

Normalmente, os documentos de texto não são armazenados por inteiro em um sistema de RI, ou seja, para cada documento são criadas estruturas de dados com o objetivo de acelerar o processo de recuperação. Os documentos que devem ser indexados são submetidos a um processo de filtragem de termos relevantes, denominado “extração de atributos”. Nesta pesquisa, todos os termos dos documentos considerados na estrutura da MO passaram pela etapa de pré-processamento de texto. A Figura 13 apresenta a síntese das etapas envolvidas na construção da base de termos.

Figura 13 – Etapas da construção da base de termos




Fonte: Elaborada pela autora

Depois de selecionar um conjunto de documentos pertencentes a uma categoria ou mais, inicia-se a fase de preparação dos documentos para construção da base de termos. Essa etapa compreende diversos passos para transformar esse conjunto de documentos (conjunto de arquivos), em linguagem natural, em uma lista de termos úteis que podem ser carregados como vetores para a base de termos.

Nessa etapa do pré-processamento dos documentos de texto, realizaram-se estudos comparativos para identificar uma forma mais rápida de implementar esse processo. Ou seja, analisou-se

comparativamente uma abordagem serializada com uma abordagem paralelizada (WILGES et al., 2013). Assim, a implementação serializada foi realizada pelo RapidMiner e, para explorar o potencial de paralelização, foi utilizada, através do paradigma de programação do MapReduce, a tecnologia do Hadoop. O que motivou esta comparação foi a exploração de um grande conjunto de documentos de texto para construção da base de termos por categoria. Os resultados dessa pesquisa apresentam uma eficiência maior no tempo de processamento em série, provavelmente porque os arquivos de texto utilizados no processamento distribuído não tinham um tamanho da ordem de petabytes, o que normalmente é sugerido para processamento de grandes massas de dados.

Dessa forma, todos os termos para cada categoria da BD foram ordenados e normalizados a partir da frequência mais alta. Assim, definiu-se um índice denominado "grau de relevância" (*relevance degree*), obtido para cada termo. O grau de relevância de cada termo deve ser calculado para todas as categorias de forma independente. Assim, o peso dos termos na BD para cada categoria, denominado grau de relevância do termo ($\mathrm{RD}_{\mathrm{w_i}}^{\mathrm{c_j}}$), foi calculado da seguinte forma:

$$RD_{w_i}^{c_j} = \frac{f_{(w_i,c_j)}}{\max(f_{(w_z,c_j)})} \qquad \text{(Equação 10)}$$

Onde:

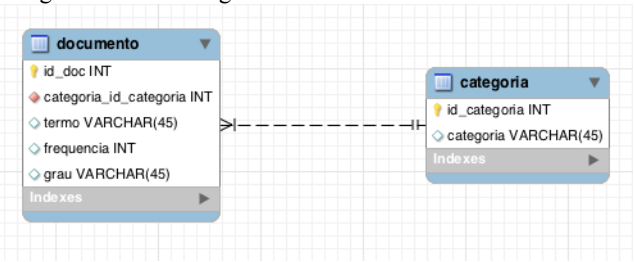
$RD_{w_i}^{c_j}$ = grau de relevância do termo $w_i$ na categoria $c_j$;

$f_{(w_i,c_j)}$ = frequência do termo $w_i$ na categoria $c_j$;

$\max(f_{(w_z,c_j)})$ = maior frequência representada pelo termo $w_z$ na categoria $c_j$.

Para cada uma das categorias analisadas no modelo de organização de documentos, a base deve armazenar uma chave para identificação do documento, uma chave para identificação da categoria, a representação do termo, a frequência do termo na categoria e o seu grau de relevância para a categoria. A Figura 14 expressa a modelagem conceitual da base de documentos.

Figura 14 – Modelagem conceitual da base de documentos




Fonte: Elaborada pela autora

O objetivo da base de termos é identificar todas as palavras que melhor expressem as características de uma determinada categoria. Essa BD é construída em função dos termos mais comuns entre os diversos documentos de uma mesma categoria. Neste caso, a BD deve ser grande o suficiente para ter uma cobertura dos termos que definem uma categoria. De acordo com Sato (2009), existem aproximadamente 400 mil palavras na língua portuguesa. Assim, se uma determinada categoria tiver pelo menos 100 mil palavras, a base projetada deverá ter um tamanho mínimo de amostra de 380 termos por categoria, considerando uma margem de erro amostral de 5%, com um nível de confiança de 95%.

Dessa forma, é possível gerar uma amostra de dados que permita por meio de comparações definir se um texto qualquer pertence a uma categoria da BD.

### 3.1.2 Definição das variáveis

Para definição das variáveis foram utilizadas funções que expressam a capacidade de comparar um documento ou uma coleção de documentos testes com os termos armazenados na base de dados. Essas funções devem retornar valores que representem a congeneridade do documento teste com as categorias armazenadas na base. Assim, cada documento teste apresentará um valor de semelhança com cada categoria considerada na base de dados em função de duas variáveis determinadas nesta pesquisa: similaridade e confiança.

A variável similaridade é definida em consonância com a variável que calcula a similaridade em RI apresentada na literatura; porém ela é adaptada, pois não trata com termos de consulta definidos pelo usuário. Em vez disso, ela analisa um documento teste qualquer em relação ao

conjunto de categorias consideradas em uma base de dados e retorna seu valor de similaridade para cada categoria.

A variável confiança é definida como um apoio aos resultados apresentados pela variável similaridade, ou seja, considera-se o quanto um documento teste é similar a uma categoria em relação ao total de termos que possuem correspondência com os termos expressos na BD.

Assim, a definição dessas variáveis é semelhante à proposta do modelo vetorial de RI porque considera os pesos associados aos termos como vetores para cada categoria. De modo geral, o modelo vetorial assume que os termos encontrados nos documentos são independentes, fato que não é verdade; porém, de acordo com a literatura, modelar as dependências entre os termos em um documento tem gerado soluções com baixo desempenho em relação ao tempo e espaço e não se tem encontrado soluções melhores que as implementações do modelo vetorial.

Assim, foram utilizadas as variáveis similaridade e confiança para construção do método de classificação dos documentos, ou seja, a similaridade $S_{PT}^{C}$ entre as partes envolvidas: o texto analisado (PT) e os textos da DB. E, por outro lado, a confiança (*accuracy*) $A_{PT}^{C}$, que mede a confiabilidade dos resultados apresentados pela variável similaridade.

O texto analisado (PT), também definido como documento teste, passou pelo mesmo pré-processamento que o conjunto de documentos que compõe a base de dados, ou seja, houve o processo de extração, limpeza, remoção de *stopwords* e transformação dos caracteres em minúsculas.

Com a frequência de cada um dos termos do texto analisado $PT_{w_i}$, realizou-se a multiplicação pelo respectivo grau de relevância do termo em cada uma das categorias da $DB_{w_i}$. A Equação 11 expressa a relação entre os termos do $PT$ e os graus de relevância dos termos na $DB$, onde $f_{PTw_i}$é a frequência dos termos no texto analisado e $RD_{w_i}^{c}$é o grau de relevância do termo para uma categoria específica da $DB$. O valor associado $V_{PTw_i}$ armazena a frequência de cada termo no $PT$ multiplicado pelo seu respectivo grau de relevância dentro da categoria (Equação 10).

$$V_{PTw_i}^{c} = f_{PTw_i} * RD_{w_i}^{c} \qquad \text{(Equação 11)}$$

Onde:
$V_{PTw_i}$ = valor associado para cada palavra no texto ($PT$);

$f_{PTw_i}$ = frequência de cada palavra no texto ($PT$);
$RD_{w_i}^{c}$= grau de relevância do termo na ($DB$).

Por meio do valor associado de cada palavra $V_{PTw}$, é extraída a média do conjunto de valores associados para o texto analisado ($AV_{PT_w}$) representado pela Equação 12.

$$AV_{PT}^{c} = \frac{1}{nw_{PT}} \sum_{i=1}^{nw_{PT}} V_{PTw_i}^{c} \qquad \text{(Equação 12)}$$

Onde:
$AV_{PT}^{c}$= valor médio associado no texto analisado ($PT$);
$nw_{PT}$= total de termos no texto analisado ($PT$);
$\frac{1}{nw_{PT}} \sum_{i=1}^{nw_{PT}} V_{PTw_i}^{c}$ = somatório dos valores associados de cada palavra no texto analisado ($PT$) dividido pelo total de termos no ($PT$).

Além do valor médio associado $AV_{PT_w}$ para cada texto analisado, é extraída a representatividade de uma categoria na $DB$, ou seja, é definido o grau médio de relevância dos termos para cada categoria na base. A Equação 13 apresenta o cálculo do grau médio de relevância dos termos na base por categoria:

$$ARD_{DB}^{c} = \frac{1}{nw_{c}} \sum_{i=1}^{nw_{c}} RD_{w_i}^{c} \qquad \text{(Equação 13)}$$

Onde:
$ARD_{DB}^{c}$= grau médio de relevância dos termos na base por categoria;
$nw_c$= número total de termos na base para uma categoria;
$\frac{1}{nw_{c}} \sum_{i=1}^{nw_{c}} RD_{w_i}^{c}$ = somatório dos graus de relevância dos termos para uma categoria na base dividido pelo total de termos dessa categoria.

Assim, a partir da definição dessas funções, é possível obter diferentes valores de similaridade e confiança para cada uma das categorias. A similaridade $S_{PT}^{c}$ do texto analisado (PT) em relação aos termos da DB é calculada de acordo com os valores médios associados ($AV_{PT}^{c}$) do texto analisado dividido pelo grau médio dos termos de cada

categoria da DB ($ARD_{DB}^{c}$). A função que expressa a variável similaridade $S_{PT}^{c}$ é obtida pela Equação 14.

$$S_{PT}^{c} = \frac{AV_{PT}^{c}}{ARD_{DB}^{c}} \qquad \text{(Equação 14)}$$

A confiança representa a quantidade de termos do texto analisado (PT) que existe correspondência com a base (DB) para cada categoria. Assim, seu cálculo é representado pela Equação 15:

$$A_{PT}^{c} = 1 - \frac{(nw_{DB}^{c} - nw_{DB \cap PT}^{c})}{nw_{PT}} \qquad \text{(Equação 15)}$$

Onde:

$nw_{DB}^{c}$: total de termos para cada categoria da DB;

$nw_{DB \cap PT}^{c}$: total de termos que são comuns entre a DB e o texto analisado (PT);

$nw_{PT}$: total de termos do texto analisado (PT).

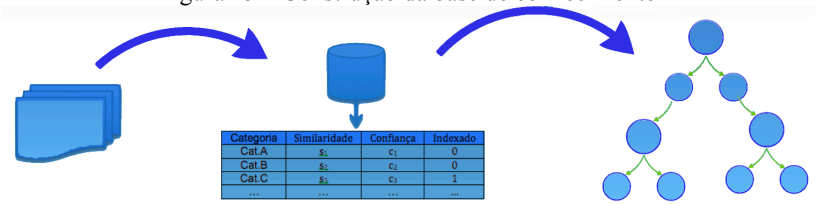
### 3.1.3 Definição da base de conhecimento

Um conjunto de documentos testes, dentro das categorias especificadas na base de termos, é selecionado para montar uma base de conhecimento sobre os resultados do cálculo das variáveis similaridade e confiança. Assim, todo esse conjunto de documentos testes iniciais passa pela etapa de pré-processamento, da mesma forma que os documentos que geraram a base de termos. Depois dessa etapa, o vetor que representa o documento teste é processado e obtém-se um valor de similaridade e outro de confiança para cada categoria considerada. Além disso, para cada documento teste é considerada sua indexação prévia a uma categoria, ou seja, sua categoria inicialmente reconhecida de onde o documento foi extraído.

Todos os resultados para o conjunto de documentos testes são armazenados em uma base de dados. Quando esse conjunto de documentos transformado em valores de similaridade e confiança é colocado em uma estrutura como um banco de dados, é possível a extração de conhecimento na base de dados (*Knowledge-Discovery in Databases*).

A Figura 15 ilustra todos os passos executados com base na coleção de documentos, desde a montagem da base de conhecimento até o processo de descoberta de conhecimento sobre a base (*Knowledge Discovery in Databases* – KDD).

Figura 15 – Construção da base de conhecimento




Fonte: Elaborada pela autora

Um processo de KDD produz conhecimento por meio do relacionamento dos dados e a sua principal característica é a extração não trivial de informações implicitamente contidas em uma base de dados. Para adquirir conhecimento na construção das regras do modelo de classificação difuso, foi fundamental o processo de KDD.

Assim, diferentes algoritmos de classificação que geram árvores de decisão (AD) são aplicados para a descoberta das relações entre as variáveis similaridade e confiança. Essa aquisição de conhecimento identifica o relacionamento entre as variáveis analisadas em diferentes situações, armazenadas na base de conhecimento, e destaca os fatores de relevância que determinam o processo de classificação de um documento.

Dessa forma, um conjunto de regras para a classificação de documentos é identificado a partir de um processo de KDD. As regras obtidas por meio da base de conhecimento são implementadas em um motor de inferência *fuzzy*. A próxima seção descreve a modelagem *fuzzy* com a utilização das variáveis similaridade e confiança e da base de regras definida.

### 3.1.4 Implementação da modelagem *fuzzy*

A lógica *fuzzy* é apropriada para tratar com fatores que relacionam ambiguidade, imprecisão e informações vagas na resolução do problema. A modelagem computacional convencional não trata com esses fatores de ambiguidade, porque utiliza apenas definições de verdadeiro ou falso. Além disso, os resultados das funções similaridade

e confiança devem considerar uma margem, mesmo que mínima, de imprecisão.

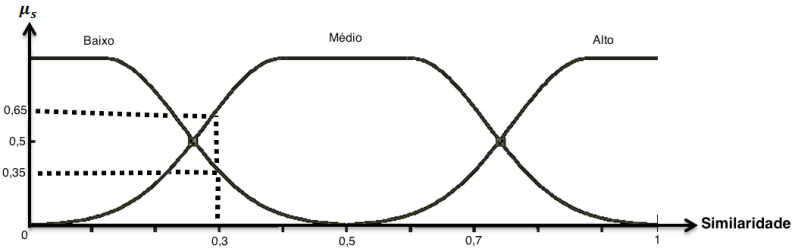
A teoria dos conjuntos *fuzzy* foi desenvolvida para tratar esses fatores de vagueza de uma forma matemática, considerando níveis de imprecisão e ambiguidade. Ao contrário da lógica tradicional, a lógica fuzzy não impõe valores rígidos e é capaz de atribuir graus de pertinência para os documentos em uma determinada categoria.

As variáveis linguísticas são elementos simbólicos utilizados para descrever o conhecimento que normalmente é reproduzido por variáveis numéricas. Essas variáveis utilizam um conjunto de termos linguísticos como grandeza de medida. Assim, o modelo proposto define a similaridade e a confiança como variáveis linguísticas de entrada, contendo os termos linguísticos: alto, médio e baixo. Esses termos associam-se às funções de pertinência por meio dos graus de pertinência e possibilitam, assim, um valor numérico. Assim as funções de pertinência das variáveis similaridade e confiança são expressas da seguinte forma:

$$\mu_s : x \rightarrow [0,1]$$
$$\mu_c : x \rightarrow [0,1] \qquad \text{(Equação 16)}$$

Nessa representação, cada elemento $x$ do universo possui um grau de pertinência tanto para a variável similaridade quanto para a variável confiança, em um intervalo fechado de 0 a 1. A Figura 16 apresenta uma representação *fuzzy* para a similaridade com suas funções de pertinência.

Figura 16 – Funções de pertinência para variável similaridade




Fonte: Elaborada pela autora

Na Figura 16 é possível observar que um valor de similaridade igual a 0,3 ativa duas funções de pertinência: baixo com $\mu_s$ = 0,35 e

médio com $\mu_s$ = 0,65. Assim, um valor de 0,3 para similaridade representa uma similaridade média, mas não deixa de ser baixa. Na lógica *fuzzy*, uma variável não é tratada como tendo apenas um estado atual, mas sim n estados, cada um com um grau de pertinência.

Considerando as duas variáveis de entrada, similaridade e confiança, e um elemento $x$ com seus respectivos de graus de pertinência para cada variável $\mu_s(x)$ e $\mu_c(x)$, as operações padrões de união, interseção e complemento são representadas respectivamente da seguinte forma:

$$\mu_{s \cup c}(x) = max[\mu_s(x), \mu_c(x)]$$
$$\mu_{s \cap c}(x) = min[\mu_s(x), \mu_c(x)] \qquad \text{(Equação 17)}$$
$$\mu_{\bar{s}}(x) = 1 - \mu_s(x)$$

Essas operações representam os procedimentos de inferência que são combinados com as regras da modelagem. A estrutura de representação que define o comportamento de uma variável *fuzzy* em relação a outra, estabelecendo um conjunto de relações condicionais, é denominada de regras do sistema. As regras nesse modelo são extraídas a partir de uma base de conhecimento. Elas são representadas por meio de condicionais do tipo *se...então* provenientes de uma estrutura de árvore de decisão.

A modelagem *fuzzy* é composta por um fuzzyficador, uma base de conhecimento, representada pelas regras, mecanismos de inferência e funções de agregação e um desfuzzyficador. A estrutura da modelagem *fuzzy* descrita é representada pela Figura 17.

Figura 17 – A estrutura da modelagem *fuzzy*


Fonte: Elaborada pela autora

Onde:

$S_{PT}^{c_n}$= valor de entrada para a variável similaridade em uma categoria $n$;

$A_{PT}^{c_n}$= valor de entrada para a variável confiança em uma categoria $n$;

$C_n =$ valor de saída para a categoria $n$.

Por meio da Figura 17 é possível observar que as entradas de similaridade e confiança, calculadas para cada categoria, passam pelo fuzzyficador para mapear seus valores em graus de pertinência aos conjuntos *fuzzy*. No centro da figura se encontra o raciocínio *fuzzy* que ativa as regras combinadas com o mecanismo de inferência. O desfuzzyficador transforma a saída *fuzzy* em uma informação numérica para cada uma das n categorias consideradas.

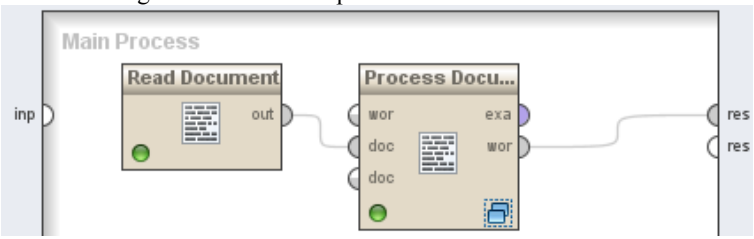
# 4 IMPLEMENTAÇÃO DO MODELO

A principal motivação para implementação desse modelo de organização foi assegurar que ele funcionasse e alçasse seus objetivos de acordo com seu propósito. Assim, na implementação inicialmente se considerou um conjunto de quatro categorias: Educação, Tecnologia, Esporte e Economia. Dessa forma, selecionou-se uma coleção de documentos nessas categorias para construir o método de classificação que apoia o modelo de organização. São estes os passos que desenvolvem a metodologia do método proposto: pré-processamento dos documentos, construção da base de termos, desenvolvimento da base de conhecimento, implementação da modelagem *fuzzy*, cujos resultados são apresentados neste capítulo.

## 4.1 PRÉ-PROCESSAMENTO DOS DOCUMENTOS

Nesta pesquisa utilizou-se uma ferramenta denominada RapidMiner (2014) para realizar tanto o pré-processamento dos documentos quanto a implementação dos algoritmos de classificação das árvores de decisão e dos *clusters*.

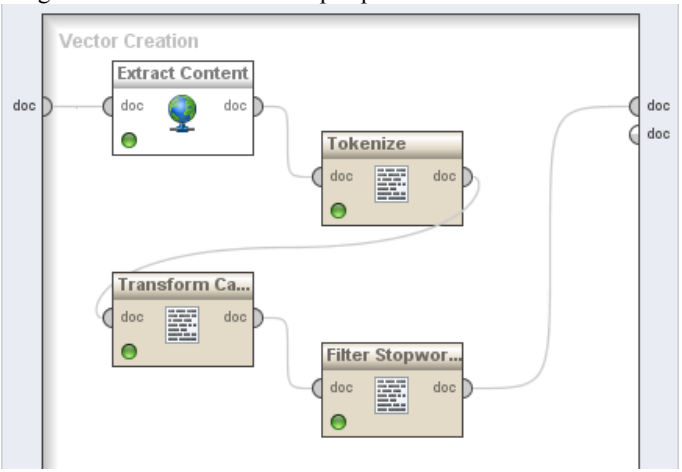
Figura 18 – Leitura e processamento do documento




Fonte: Elaborada pela autora

Assim, todos os arquivos, tanto da base de documentos quanto os documentos de teste, passaram pelas etapas de extração, limpeza, remoção de *stopwords* e transformação dos caracteres em minúsculas. Visualizam-se os detalhes do pré-processamento implementado pelo RapidMiner na Figura 19.

Figura 19 – Detalhamento do pré-processamento do documento




Fonte: Elaborada pela autora

O pré-processamento dos documentos (Figura 19) envolve primeiramente o operador Extract content, que extrai o conteúdo textual de um documento em formato HTML e retorna os blocos de texto extraídos desse documento. Posteriormente, é aplicado o operador Tokenize, que divide o texto de um documento em uma sequência de *tokens*. O próximo operador aplicado é o Transform cases, o qual transforma todos os caracteres de um documento em minúsculo ou maiúsculo. E, finalmente, o operador Filter stopwords (*dictionary*), que retira todas as palavras que pertencem a uma lista de *stopwords*, que é carregada a partir de um arquivo no próprio operador. Essa lista de *stopwords* contém as palavras que não são consideradas relevantes para a contagem dos termos mais frequentes de um documento de texto, ou seja, são palavras que não influenciam na definição de uma categoria, tais como: artigos, preposições e algumas conjunções.

A lista de *stopwords* utilizada nesse processamento foi extraída da Intext mining (2013), que disponibiliza, além desta lista de *stopwords* em português, listas em inglês e espanhol.

## 4.2 CONSTRUÇÃO DA BASE DE TERMOS

No desenvolvimento do método de classificação, considerou-se um conjunto de dados iniciais selecionados de documentos pertencentes

a quatro categorias. Assim, construiu-se a base de termos com palavras para cada uma das categorias utilizando textos não estruturados extraídos de fontes da *web*, principalmente de jornais e revistas *on-line*. Na organização dessa base de dados (BD), relacionaram-se as categorias: Educação, Tecnologia, Esporte e Economia.

Na construção da base de termos, considerou-se um total de aproximadamente 4.300 termos, observando que um tamanho mínimo de amostra por categoria deveria conter pelo menos 384 termos. Os textos foram extraídos da *web* considerando sua indexação prévia por categoria e pela facilidade de encontrar documentos para cada categoria de teste. Assim, essa base indexou 979 termos para economia, 1.210 para educação, 1.078 para esporte e 1.092 para tecnologia. A Figura 20 apresenta a tabela de categorias e a Figura 21 apresenta uma síntese da tabela de termos por categoria.

Figura 20 – Tabela de categorias

| id_categoria | categoria |
|---|---|
| 1 | Economia |
| 2 | Educação |
| 3 | Esporte |
| 4 | Tecnologia |

Fonte: Elaborada pela autora

Figura 21 – Tabela de termos por categoria

| id_doc | categoria_id_categoria | termo | frequencia | grau | valorAssociado |
|---|---|---|---|---|---|
| 1 | 1 | dólar | 15 | 1 | 15 |
| 2 | 1 | ano | 13 | 0.867 | 11.26 |
| 3 | 1 | alta | 9 | 0.6 | 5.4 |
| 4 | 1 | gás | 11 | 0.73 | 8.06 |
| 5 | 1 | us | 9 | 0.6 | 5.4 |
| 6 | 1 | mercado | 8 | 0.53 | 4.26 |

Fonte: Elaborada pela autora

Os resultados da tabela de termos por categoria (Figura 21) são parciais e sua estrutura física está implementada em um banco de dados MySQL. Apresentou-se sua definição conceitual no terceiro capítulo, sobre a especificação do modelo na Figura 14.

## 4.3 DESENVOLVIMENTO DA BASE DE CONHECIMENTO

Com o objetivo de identificar um conjunto de regras iniciais para a modelagem *fuzzy* e, também, o uso dos indicadores de similaridade e confiança definidos no terceiro capítulo, foi elaborada uma base de conhecimento, como estudo de caso, para as mesmas categorias da base de termos.

A partir dos dados da base de conhecimento, é possível reconhecer uma estrutura de regras para a modelagem *fuzzy* por meio de um processo de KDD. Assim, as regras do modelo difuso são definidas com a tarefa de mineração de dados em um conjunto de treinamento, obtido por meio do cálculo de similaridade e confiança para uma coleção de documentos; ou seja, as regras não são determinadas pelo conhecimento do especialista, mas sim de um processo de descoberta de conhecimento.

Além disso, os valores obtidos para as variáveis similaridade e confiança permitem verificar o quanto as funções que expressam essas variáveis estão condizentes com seus propósitos. Portanto, cada texto processado tem um valor de similaridade ($S_{PT}^{C}$) e um valor de confiança ($A_{PT}^{C}$) em relação a cada uma das categorias da base de termos. As Tabelas 2 a 9 apresentam uma amostra parcial de oito documentos testes com os resultados das variáveis similaridade e confiança.

O atributo "indexado", utilizado na tabela corresponde à classificação real do texto em sua fonte original de onde o documento foi extraído. Desse modo, um texto que pertence, por exemplo, a uma categoria "economia", terá uma similaridade e uma confiança atribuídas a cada uma das quatro categorias previamente definidas na base.

Tabela 2 – Resultado das funções para um texto indexado em economia

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 2 | Economia | 1 | 1 | sim |
| 2 | Educação | 0,04 | 0,09 | não |
| 2 | Esporte | 0 | 0 | não |
| 2 | Tecnologia | 0,1 | 0,09 | não |

Fonte: Elaborada pela autora

Tabela 3 – Resultado das funções para um segundo texto de economia

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 3 | Economia | 1 | 1 | sim |
| 3 | Educação | 0,08 | 0,8 | não |
| 3 | Esporte | 0 | 0 | não |
| 3 | Tecnologia | 0,34 | 0,85 | não |

Fonte: Elaborada pela autora

Tabela 4 – Resultado das funções para um texto indexado em educação

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 7 | Economia | 0 | 0 | não |
| 7 | Educação | 1 | 1 | sim |
| 7 | Esporte | 0,27 | 0,21 | não |
| 7 | Tecnologia | 0,58 | 1 | não |

Fonte: Elaborada pela autora

Tabela 5 – Resultado das funções para um segundo texto de educação

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 10 | Economia | 0,32 | 0 | não |
| 10 | Educação | 1 | 1 | sim |
| 10 | Esporte | 0 | 0,33 | não |
| 10 | Tecnologia | 0,81 | 0,96 | não |

Fonte: Elaborada pela autora

Tabela 6 – Resultado das funções para um texto indexado em esporte

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 13 | Economia | 0 | 0 | não |
| 13 | Educação | 0,06 | 0,24 | não |
| 13 | Esporte | 1 | 1 | sim |
| 13 | Tecnologia | 0,11 | 0,17 | não |

Fonte: Elaborada pela autora

Tabela 7 – Resultado das funções para um segundo texto de esporte

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 15 | Economia | 0,18 | 0 | não |
| 15 | Educação | 0,23 | 0,21 | não |
| 15 | Esporte | 1 | 1 | sim |
| 15 | Tecnologia | 0 | 0,03 | não |

Fonte: Elaborada pela autora

Tabela 8 – Resultado das funções para um texto indexado em tecnologia

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 21 | Economia | 0,91 | 0,29 | não |
| 21 | Educação | 0,22 | 1 | não |
| 21 | Esporte | 0 | 0 | não |
| 21 | Tecnologia | 1 | 0,86 | sim |

Fonte: Elaborada pela autora

Tabela 9 – Resultado das funções para um segundo texto de tecnologia

| Texto | Categoria | Similaridade | Confiança | Indexado |
|---|---|---|---|---|
| 24 | Economia | 0,11 | 0 | não |
| 24 | Educação | 0,44 | 0,88 | não |
| 24 | Esporte | 0 | 0,13 | não |
| 24 | Tecnologia | 1 | 1 | sim |

Fonte: Elaborada pela autora

Normalizaram-se dentro de cada categoria os resultados das variáveis similaridade e confiança apresentados nas Tabelas 2 a 9, para que ficassem representados em um intervalo de [0,1]. Esse conjunto de

resultados para as variáveis similaridade e confiança compõe uma base de conhecimento com 97 documentos. A Tabela 10 apresenta a síntese dos resultados em relação à indexação prévia do documento.

Tabela 10 – Avaliação do cálculo de similaridade e confiança

| Categoria indexada | Similaridade | Confiança |
|---|---|---|
| Sim | 93 | 85 |
| Não | 4 | 12 |
| Total | 97 | 97 |

Fonte: Elaborada pela autora

Em uma análise geral percebeu-se que os resultados de similaridade respondem corretamente à categoria de maior relevância em 96% das vezes, e à função da confiança 88% das vezes. A variável confiança tem como principal objetivo apoiar os indicadores apresentados pela variável similaridade em relação à quantidade de termos do texto que correspondem à base de termos.

Esses resultados mostram que as definições das variáveis similaridade e confiança estão adequadas para o método do classificador do modelo de organização. Além disso, considerando um tamanho de amostra real com 10.000 documentos em uma base de conhecimento, o tamanho mínimo da amostra deveria ser 95 documentos com um nível de confiança de 95%, e uma margem de erro de 10%. A base de conhecimento elaborada neste trabalho possui 97 documentos, totalizando 388 resultados, porque cada documento é classificado em quatro categorias inicialmente definidas. De acordo com o valor de amostra apresentado, essa base é suficiente para produzir resultados em um processo de descoberta de conhecimento.

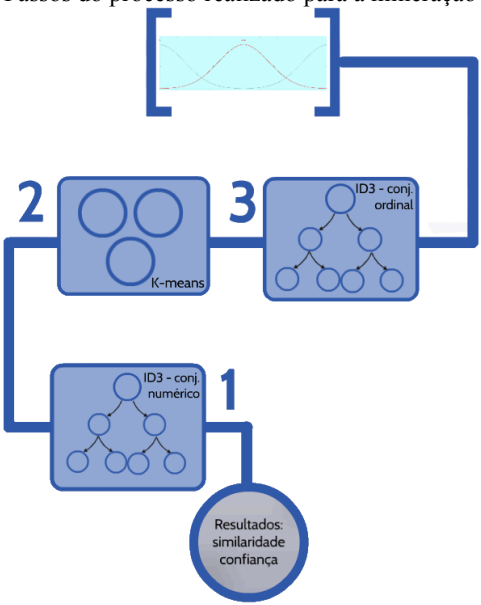
### 4.3.1 Processo de descoberta de conhecimento (KDD)

Para reconhecer padrões de dados da base de conhecimento, viabilizando as definições e o comportamento das regras da modelagem *fuzzy*, consideraram-se diferentes passos para a extração do conhecimento. Esses passos envolveram a mineração de dados para classificação com a abordagem simbólica e descritiva. Embora tenha sido fundamental a utilização da abordagem descritiva para definição dos grupos (clusters), este trabalho concentrou esforços na aplicação da abordagem simbólica para processá-los.

Na abordagem descritiva utilizou-se o algoritmo de *K-means* para implementação dos *clusters*, e na abordagem simbólica utilizou-se o algoritmo ID3, o qual implementa as árvores de decisão. O uso desses algoritmos contribui nos ajustes das funções de pertinência e principalmente nas definições das regras para a modelagem *fuzzy*. Dessa forma, pode-se aumentar a expressividade das funções de pertinência para obter resultados mais condizentes com os valores esperados no processo de classificação, sem a necessidade de definições manuais desses parâmetros.

Assim, utilizou-se a base de conhecimento como um conjunto de aprendizado, que foi dividido de acordo com o método *holdout* em um conjunto de treinamento de 2/3 e o conjunto de teste 1/3. Dessa maneira, utilizou-se um conjunto de treinamento com 65 documentos, totalizando 260 resultados, e um conjunto de teste com 32 documentos, totalizando 128 resultados. A Figura 22 apresenta os passos envolvidos no processo de mineração dos dados.

Figura 22 – Passos do processo realizado para a mineração de dados

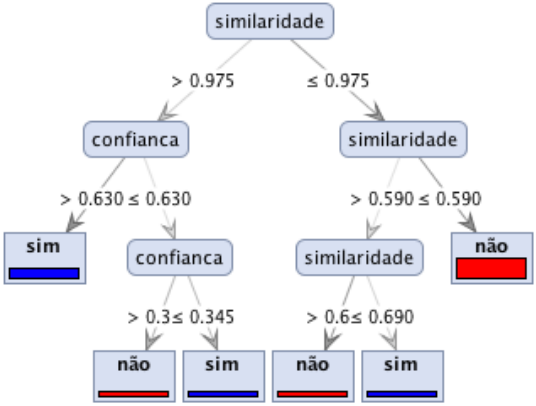



Fonte: Elaborada pela autora

O primeiro passo foi a aplicação do algoritmo ID3 diretamente ao conjunto de treinamento com os valores de similaridade e confiança já

calculados. A Figura 23 mostra a árvore gerada pelo algoritmo, com os valores para similaridade e confiança relacionados com o *label* indexado (sim ou não) para todas as categorias consideradas.

Figura 23 – Árvore gerada pelo algoritmo ID3 numérico




Fonte: Elaborada pela autora

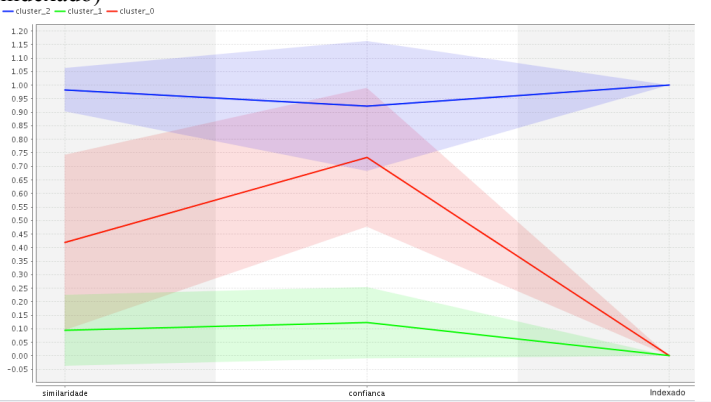
A escolha central do algoritmo ID3 está em sua capacidade de selecionar qual atributo é mais útil para a construção da AD, ou seja, utiliza-se o conceito de ganho de informação, que mede como um determinado atributo separa os exemplos de treinamento de acordo com a classificação destes. O conceito de ganho de informação é relacionado com a entropia e caracteriza a impureza de uma coleção arbitrária. Dessa forma, quanto menor o valor da entropia, menor é a incerteza e mais útil é o atributo para a classificação.

Na estrutura que se mostra na Figura 23, o atributo que é identificado como o menor valor de entropia é a similaridade. No algoritmo ID3, cada vértice ou nodo corresponde a um atributo, e cada aresta da árvore, a um valor possível do atributo. Cada vértice é posicionado na AD segundo seu nível de informação, calculado por meio do conceito de entropia. Assim, as folhas da árvore correspondem ao valor esperado da decisão segundo os dados de treinamento utilizados.

Com esse primeiro passo (Figura 22) e objetivando um detalhamento mais apurado para a construção das regras da modelagem *fuzzy*, visualizou-se a necessidade de agrupar os conjuntos numéricos das variáveis similaridade e confiança em conjuntos ordinais.

Mas, antes disso, analisou-se a distribuição do conjunto de treinamento com os valores de similaridade e confiança em relação a sua indexação à categoria original. Na Figura 24 é possível observar, nesse agrupamento geral dos dados do conjunto de treinamento, que quando os valores da similaridade e confiança são altos, existe uma conversão para a classe indexado com valor 1. Observa-se, também, que quando os valores da variável confiança são medianos a altos, mas a variável similaridade é mediana, não há conversão para a classe indexado com valor 1.

Figura 24 – Agrupamento do conjunto de teste (similaridade, confiança e indexado)




Fonte: Elaborada pela autora

Os valores numéricos da AD gerada no passo 1 foram *clusterizados* em conjuntos: baixo, médio e alto (passo 2 da Figura 22). A Figura 25 apresenta parte da base utilizada nesse processo de agrupamento aplicando o algoritmo *K-means*.

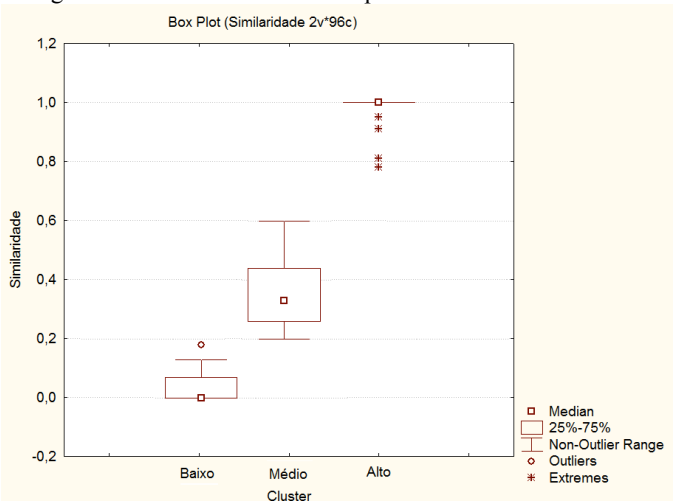
Figura 25 – Parte da base utilizada no processo de agrupamento

| id | cluster | similaridade | confianca | Indexado |
|---|---|---|---|---|
| 1 | cluster_0 | 0.360 | 0.510 | 1 |
| 2 | cluster_0 | 0.620 | 0.490 | 1 |
| 3 | cluster_0 | 0.290 | 0.450 | 1 |
| 4 | cluster_0 | 0.310 | 0.440 | 1 |
| 5 | cluster_0 | 0.490 | 0.400 | 1 |
| 6 | cluster_0 | 0.190 | 0.370 | 1 |
| 10 | cluster_0 | 0.190 | 0.270 | 1 |
| 12 | cluster_0 | 0.180 | 0.270 | 1 |
| 7 | cluster_1 | 0.170 | 0.360 | 0 |
| 8 | cluster_1 | 0.160 | 0.290 | 0 |
| 9 | cluster_1 | 0.100 | 0.280 | 0 |
| 11 | cluster_1 | 0.130 | 0.270 | 0 |
| 13 | cluster_1 | 0.150 | 0.260 | 0 |
| 14 | cluster_1 | 0.120 | 0.260 | 0 |
| 15 | cluster_1 | 0.180 | 0.250 | 0 |
| 16 | cluster_1 | 0.110 | 0.250 | 0 |
| 17 | cluster_1 | 0.100 | 0.250 | 0 |
| 18 | cluster_1 | 0.150 | 0.240 | 0 |
| 19 | cluster_1 | 0.140 | 0.240 | 0 |
| 20 | cluster_1 | 0.110 | 0.240 | 0 |

Fonte: Elaborada pela autora

O algoritmo *K-means* foi utilizado para gerar os três conjuntos ordinais (baixo, médio e alto) para as variáveis similaridade e confiança. Apresenta-se na Figura 26 o resultado do agrupamento para a variável similaridade. Por meio desse processo de agrupamento foi possível realizar uma triagem identificando *outliers* na amostra.

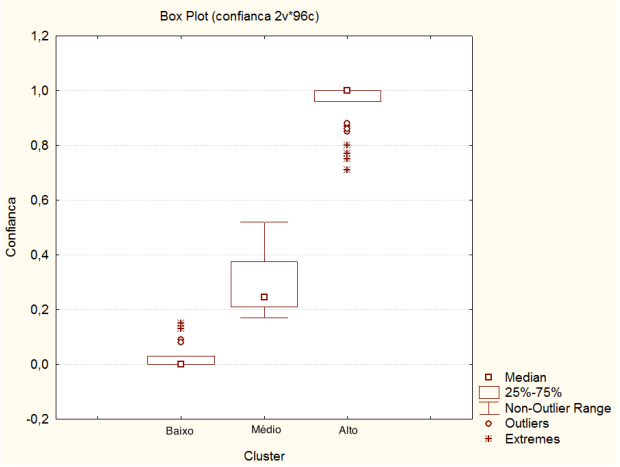
Figura 26 – Divisão dos *clusters* para a variável similaridade




Fonte: Elaborada pela autora

Na Figura 26 o *cluster* baixo está compreendido entre: 0 e 0,18. O *cluster* médio está compreendido entre: 0,20 e 0,60 e o *cluster* alto está compreendido entre: 0,78 e 1. A Figura 27 apresenta os resultados gerados para o agrupamento da variável confiança com algoritmo *K-means*.

Figura 27 – Divisão dos *clusters* da variável confiança

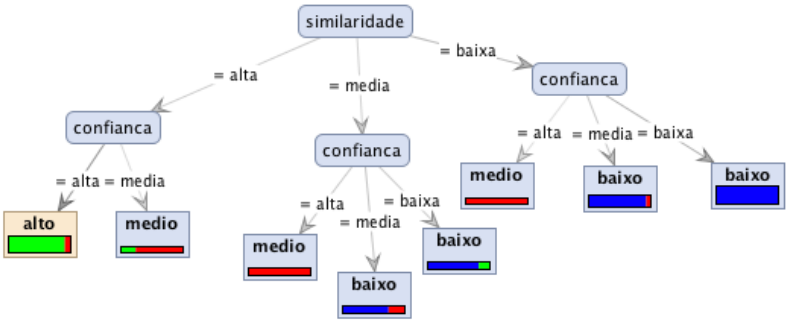



Fonte: Elaborada pela autora

Na Figura 27 o *cluster* baixo está compreendido entre 0 e 0,15, o *cluster* médio está compreendido entre 0,17 e 0,52, e o *cluster* alto está compreendido entre 0,71 e 1.

Com esses agrupamentos houve uma transformação dos dados do conjunto de treinamento para uma codificação simbólica, ou seja, os dados quantitativos foram transformados em qualitativos. O conjunto de treinamento transformado foi novamente processado pelo algoritmo ID3 gerando a AD da Figura 28.

Figura 28 – AD gerada pelo algoritmo ID3 com base no agrupamento




Fonte: Elaborada pela autora

Observa-se, na AD da Figura 28, que quando a similaridade é alta ($S_A$) e a confiança é alta ($C_A$), o texto pertence ($P_S$) a uma categoria com um grau de relevância significativo (alto), em 91% dos casos. Quando a similaridade é média ($S_M$) e a confiança é alta ($C_A$), o texto pertence ($P_S$) com um grau de relevância médio à categoria, em 100% dos casos. Quando a similaridade é média ($S_M$) e a confiança é média ($C_M$) ou baixa ($C_B$), o texto pertence à categoria com um grau de relevância baixo. Nos casos em que a similaridade possui um valor baixo ($S_B$), se a confiança é alta ($C_A$), o texto pertence à categoria com um grau médio, mas se a confiança é média ($C_M$) ou baixa ($C_B$), o texto tem uma aderência baixa à categoria.

Com o conjunto de testes de 32 documentos, foram realizadas as verificações do modelo da AD (Figura 28), AD gerada pelo processamento do conjunto de treinamento. Os resultados apresentados pela matriz do Quadro 7 avaliam o desempenho do modelo da AD.

Quadro 7 – Matriz de avaliação do desempenho do modelo da AD

| **Matriz de correspondência** | | **Classe prevista** | |
|---|---|---|---|
| | | Sim | Não |
| **Classe real** | Sim | 26 | 6 |
| | Não | 8 | 88 |

Fonte: Elaborado pela autora

A acurácia e o erro do modelo consideram os resultados das correspondências entre a classe real e a classe prevista observados no Quadro 7, e seguiram as definições dessas funções conforme apresentadas na metodologia deste trabalho. Assim, o valor resultante da acurácia do modelo foi 89%. Considerou-se adequado o modelo proposto pela AD da Figura 27 para estimar as regras da modelagem difusa, porque ele possui um valor de acurácia considerado positivo mesmo que a margem de erro não seja tão baixa.

O maior desafio no desenvolvimento da modelagem *fuzzy* foi simular a base de regras de forma que ela espelhasse o conhecimento impresso pela AD. A implementação de uma base regras nesse formato é interessante porque considera um comportamento padrão para o relacionamento e disposição das variáveis similaridade e confiança.

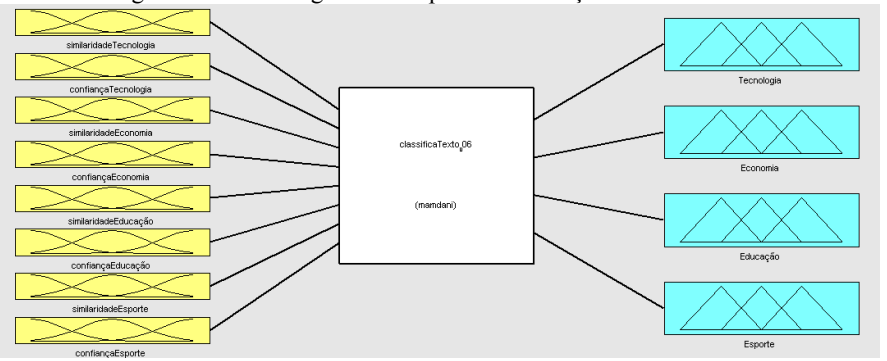
## 4.4 DESENVOLVIMENTO DA MODELAGEM *FUZZY*

A modelagem *fuzzy* para classificação de textos em múltiplas categorias é pertinente porque consegue lidar com a sobreposição de categorias que podem classificar um documento. Assim, as variáveis de entrada do sistema são tratadas por termos linguísticos que consideram essa imprecisão. Além disso, a lógica *fuzzy* é capaz de sintetizar os resultados de uma classificação em regras, considerando duas entradas (similaridade e confiança) em uma saída numérica desfuzzyficada.

Dessa forma, de acordo com a estrutura da AD ID3, tanto a variável de entrada similaridade quanto a variável confiança foram definidas com as funções de pertinência: alta, média e baixa. Cada variável de entrada foi projetada para considerar seu valor de similaridade e confiança por categoria. Além do mais, como a proposta desta pesquisa foi desenhada para responder a múltiplas categorias, então as variáveis de saída foram replicadas para responder a cada uma

das classes envolvidas: Tecnologia, Educação, Esporte e Economia. A Figura 29 mostra uma visão geral do modelo difuso implementado.

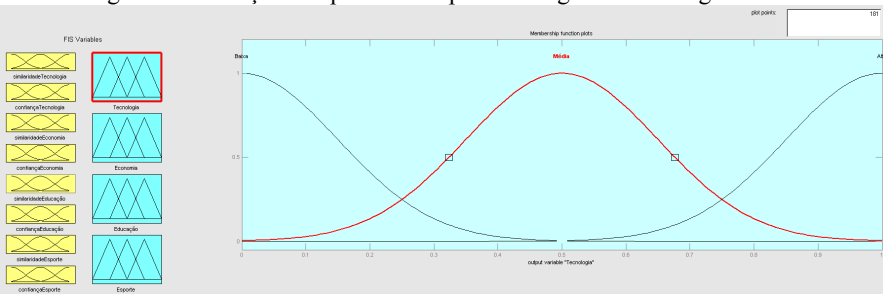
Figura 29 – Modelagem difusa para classificação de textos




Fonte: Elaborada pela autora

As funções de pertinência foram desenvolvidas com a distribuição de Gauss, utilizando a função *gaussmf*. Essa função tem um formato de curva de sino com máximo em 1 e mínimo em 0. A Figura 30 ilustra as funções de pertinência para uma das categorias.

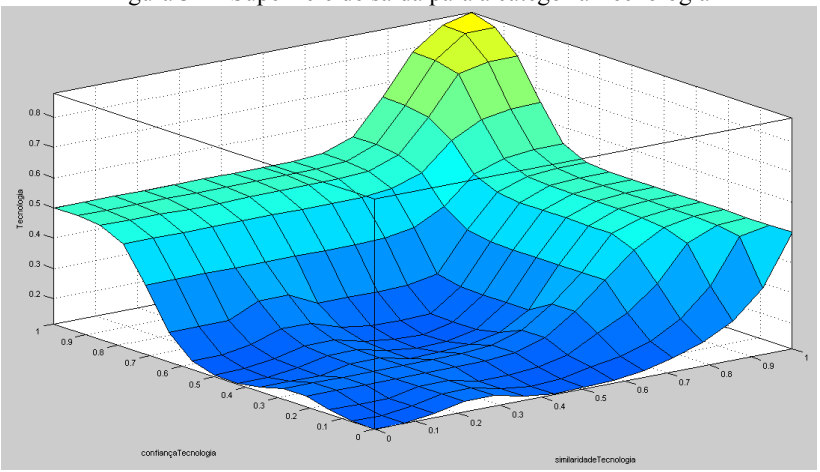
Figura 30 – Funções de pertinência para a categoria Tecnologia




Fonte: Elaborada pela autora

Além de a distribuição da função *gaussmf* ter uma curva suave, ela responde melhor nos testes de classificação dos textos porque é capaz de suavizar a superfície de saída (Figura 31).

Figura 31 – Superfície de saída para a categoria Tecnologia




Fonte: Elaborada pela autora

Nesta modelagem difusa o controlador do motor de inferência aplicado foi *Mamdani*, implementado com o operador mínimo e o método desfuzzyficação foi o centro de massa. As regras de inferência são apresentadas na Figura 32 e foram construídas em consonância com as definições apresentadas no resultado da AD gerada pelo algoritmo ID3 (Figura 28). As funções de pertinência do modelo difuso foram ajustadas até que a saída gerada pelo modelo se aproximasse dos valores apresentados pela AD.

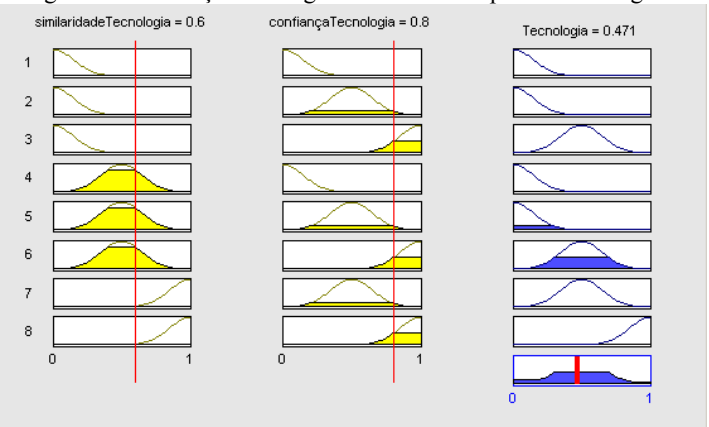
Figura 32 – Base de regras do modelo difuso

1. If (similaridadeTecnologia is Baixa) and (confiançaTecnologia is Baixa) then (Tecnologia is Baixa) (1)
2. If (similaridadeTecnologia is Baixa) and (confiançaTecnologia is Média) then (Tecnologia is Baixa) (1)
3. If (similaridadeTecnologia is Baixa) and (confiançaTecnologia is Alta) then (Tecnologia is Média) (1)
4. If (similaridadeTecnologia is Média) and (confiançaTecnologia is Baixa) then (Tecnologia is Baixa) (1)
5. If (similaridadeTecnologia is Média) and (confiançaTecnologia is Média) then (Tecnologia is Baixa) (1)
6. If (similaridadeTecnologia is Média) and (confiançaTecnologia is Alta) then (Tecnologia is Média) (1)
7. If (similaridadeTecnologia is Alta) and (confiançaTecnologia is Média) then (Tecnologia is Média) (1)
8. If (similaridadeTecnologia is Alta) and (confiançaTecnologia is Alta) then (Tecnologia is Alta) (1)
9. If (similaridadeEconomia is Baixa) and (confiançaEconomia is Baixa) then (Economia is Baixa) (1)
10. If (similaridadeEconomia is Baixa) and (confiançaEconomia is Média) then (Economia is Baixa) (1)
11. If (similaridadeEconomia is Baixa) and (confiançaEconomia is Alta) then (Economia is Média) (1)
12. If (similaridadeEconomia is Média) and (confiançaEconomia is Baixa) then (Economia is Baixa) (1)
13. If (similaridadeEconomia is Média) and (confiançaEconomia is Média) then (Economia is Baixa) (1)
14. If (similaridadeEconomia is Média) and (confiançaEconomia is Alta) then (Economia is Média) (1)
15. If (similaridadeEconomia is Alta) and (confiançaEconomia is Média) then (Economia is Média) (1)
16. If (similaridadeEconomia is Alta) and (confiançaEconomia is Alta) then (Economia is Alta) (1)

Fonte: Elaborada pela autora

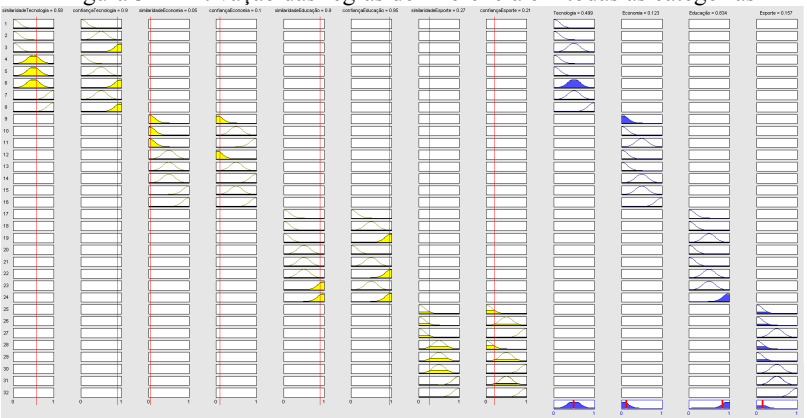
Foram definidas oito regras de inferência para cada categoria, que cobrem todas as variáveis linguísticas abordadas na AD. Cada regra consiste no operador AND associado ao método mínimo. A agregação das regras é feita por meio do método máximo. Os resultados da modelagem difusa são relevantes porque podem permitir a ativação das funções de pertinência em até dois pontos para cada uma das variáveis de entrada. A Figura 33 apresenta a ativação das regras de inferência, em uma visão simplificada, para uma das categorias. Obtém-se a visão completa na Figura 34.

Figura 33 – Ativação das regras de inferência para uma categoria




Fonte: Elaborada pela autora

Figura 34 – Ativação das regras de inferência em todas as categorias



Fonte: Elaborada pela autora

### 4.4.1 Resultados da modelagem *fuzzy*

Sabe-se que, normalmente, existe uma categoria predominante ao realizar o processo de classificação do documento. Ainda assim, existem outros valores de relevância associados ao documento em relação a cada uma das categorias consideradas no domínio. Esses valores variam em um intervalo de [0,1].

Os resultados do modelo de organização de documento, por meio do classificador difuso, com suas respectivas entradas para similaridade e confiança em cada uma das categorias são apresentados na Tabela 11. Esses resultados apresentam em síntese a saída desfuzzyficada da modelagem *fuzzy* para cada uma das entradas.

Tabela 11 – Resultado do classificador *fuzzy* em múltiplas categorias

| Texto | Categoria | Sim. | Conf. | Index | Saída *fuzzy* |
|---|---|---|---|---|---|
| 2 | Economia | 1 | 1 | sim | 0,877 |
| 2 | Educação | 0,04 | 0,09 | não | 0,121 |
| 2 | Esporte | 0 | 0 | não | 0,265 |
| 2 | Tecnologia | 0,1 | 0,09 | não | 0,123 |
| 3 | Economia | 1 | 1 | sim | 0,877 |
| 3 | Educação | 0,08 | 0,8 | não | 0,467 |
| 3 | Esporte | 0 | 0 | não | 0,117 |
| 3 | Tecnologia | 0,34 | 0,85 | não | 0,490 |
| 7 | Economia | 0 | 0 | não | 0,117 |
| 7 | Educação | 1 | 1 | sim | 0,877 |
| 7 | Esporte | 0,27 | 0,21 | não | 0,157 |
| 7 | Tecnologia | 0,58 | 1 | não | 0,501 |
| 10 | Economia | 0,32 | 0 | não | 0,141 |
| 10 | Educação | 1 | 1 | sim | 0,877 |
| 10 | Esporte | 0 | 0,33 | não | 0,139 |
| 10 | Tecnologia | 0,81 | 0,96 | não | 0,703 |
| 13 | Economia | 0 | 0 | não | 0,117 |
| 13 | Educação | 0,06 | 0,24 | não | 0,161 |
| 13 | Esporte | 1 | 1 | sim | 0,877 |
| 13 | Tecnologia | 0,11 | 0,17 | não | 0,138 |
| 15 | Economia | 0,18 | 0 | não | 0,141 |
| 15 | Educação | 0,23 | 0,21 | não | 0,157 |
| 15 | Esporte | 1 | 1 | sim | 0,877 |
| 15 | Tecnologia | 0 | 0,03 | não | 0,117 |
| 21 | Economia | 0,91 | 0,29 | não | 0,500 |
| 21 | Educação | 0,22 | 1 | não | 0,497 |
| 21 | Esporte | 0 | 0 | não | 0,117 |
| 21 | Tecnologia | 1 | 0,86 | sim | 0,790 |
| 24 | Economia | 0,11 | 0 | não | 0,125 |
| 24 | Educação | 0,44 | 0,88 | não | 0,496 |
| 24 | Esporte | 0 | 0,13 | não | 0,129 |
| 24 | Tecnologia | 1 | 1 | sim | 0,877 |

Fonte: Elaborada pela autora

A avaliação geral desses resultados foi realizada em relação à atribuição do valor da saída desfuzzyficada para a categoria de indexação original do documento. Em alguns casos, percebe-se que, mesmo que a similaridade seja relativamente alta, se a confiança não representar um valor significativo para uma mesma categoria, o modelo difuso prioriza o resultado do par (similaridade e confiança) mais significativo no conjunto de resultados.

A Tabela 12 apresenta a síntese dos resultados gerais, considerando o conjunto de 97 documentos, em relação à indexação prévia do documento a uma determinada categorias e à saída *fuzzy* gerada pela modelagem.

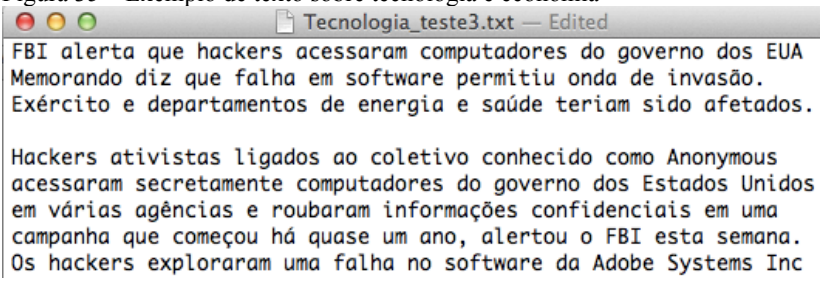
Tabela 12 – Avaliação da saída desfuzzyficada

| **Categoria indexada** | **Escore** |
|---|---|
| Sim | 76 |
| Não | 21 |
| Total | 97 |

Fonte: Elaborada pela autora

De acordo com a Tabela 12, a taxa de acerto nos 97 documentos do conjunto de treinamento classificados em relação à indexação original e o valor da saída desfuzzyficado foi 78%. Observou-se, também, que os textos tinham pertinência a outros conjuntos, como o texto 21 apresentado na Tabela 11. A Figura 35, por exemplo, apresenta parte do anúncio da notícia desse documento indexado como Tecnologia.

Figura 35 – Exemplo de texto sobre tecnologia e economia




Fonte: Elaborada pela autora

Nesse documento, identifica-se uma saída *fuzzy* alta também para a categoria economia. Na Tabela 11, a categoria economia é a segunda

mais relevante para o documento. De fato, esse texto trata do acesso a computadores do governo por *hackers*. E a quantidade de termos que se referem ao governo nesse documento é significativa para classificá-lo, além da categoria de tecnologia, na categoria economia.

# 5 AVALIAÇÃO DO MODELO

A aplicação do modelo de organização de documentos objetiva avaliar se um documento ou uma coleção de documento contém características de interesse para a organização. Assim, o espaço de busca é reduzido e os documentos podem ser estruturados em coleções dentro da MO. Além do mais, diferentes indivíduos da organização podem compartilhar documentos ou informações que são de interesse comum.

## 5.1 A ORGANIZAÇÃO PROPOSTA

Os resultados deste capítulo referem-se à avaliação do modelo no contexto de uma organização de Tecnologia da Informação (TI) com uma MO. A ontologia de domínio dessa organização focaliza um conjunto com 17 categorias predefinidas e de importância para a organização. A Figura 36 ilustra uma ontologia para essa organização com base no modelo apresentado no segundo capítulo (Figura 4). Essa ontologia representa o Domínio, a Informação e a Organização como classes. As classes Domínio e Organização são disjuntas por não compartilharem as mesmas instâncias, diferentemente das classes Domínio e Informação, que relacionam subclasses e atributos.

Figura 36 – Ontologia da organização de TI


Fonte: Elaborada pela autora

Para que a estrutura dessa ontologia tenha o comportamento especificado, é necessária a definição de uma restrição que permita que a propriedade "possuir" relacione a classe Informação com a subclasse Conteúdo da classe Domínio. Assim, é possível relacionar um documento ao conteúdo conforme a descrição da ontologia (Figura 38).

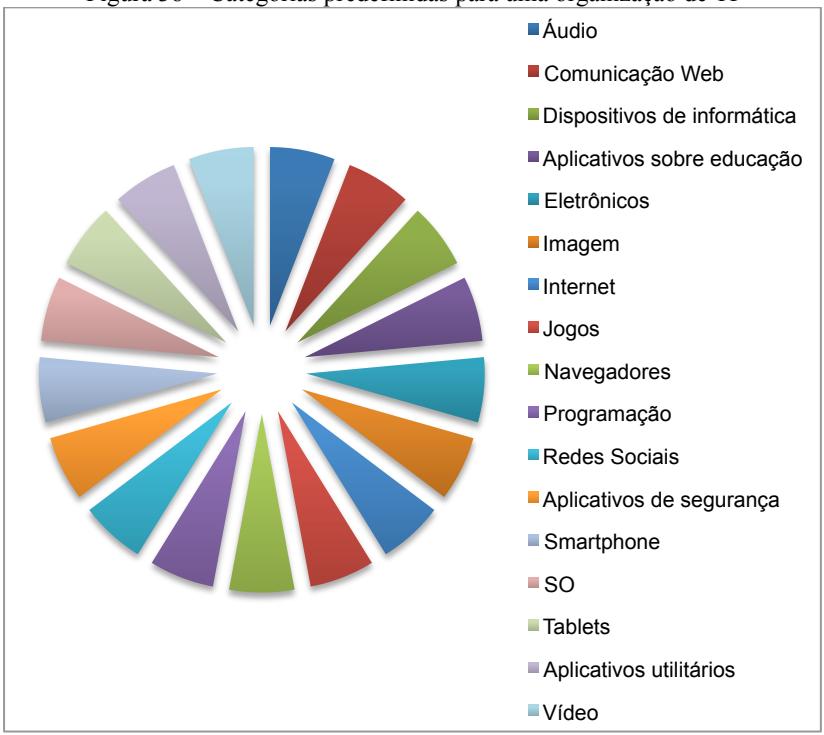
Figura 37 – Aplicação de restrições na ontologia


Fonte: Elaborada pela autora

As ontologias possuem diferentes classificações que variam de acordo com seu grau de genericidade. Elas podem ser classificadas como ontologia de representação, geral, central, de domínio, e de aplicação. Esta proposta mescla uma ontologia de representação com uma ontologia de domínio.

Dessa forma, foram utilizados conteúdos que indicassem informações sobre categorias de interesse para essa organização de TI. O resultado foi uma coleção com as seguintes categorias: áudio, comunicação *web*, dispositivos de informática, aplicativos sobre educação, eletrônicos, imagem, internet, jogos, navegadores, programação, redes sociais, aplicativos de segurança, *smartphone*, sistemas operacionais (SO), *tablets*, aplicativos utilitários e vídeos. A Figura 38 apresenta todas essas categorias da organização de TI proposta.

Figura 38 – Categorias predefinidas para uma organização de TI




Fonte: Elaborada pela autora

Os documentos necessários para a construção dessa base de termos da organização de TI bem como os documentos para testar o classificador difuso foram provenientes dos canais *web*:

- http://www.techtudo.com.br
- http://olhardigital.uol.com.br
- http://tecnologia.terra.com.br/
- http://tecnologia.uol.com.br/
- http://g1.globo.com/tecnologia/
- http://exame.abril.com.br/topicos/tecnologia

No total foram coletados cerca de 140 documentos, totalizando aproximadamente 16.000 termos para todas as categorias especificadas. A Tabela 14 expressa o total de termos utilizados para representar cada categoria dessa base no modelo.

Tabela 13 – Total de termos por categoria na organização de TI

| **Categoria** | **total** |
|---|---|
| Áudio | 821 |
| Comunicação *web* | 637 |
| Dispositivos de informática | 1637 |
| Aplicativos sobre educação | 863 |
| Eletrônicos | 624 |
| Imagem | 808 |
| Internet | 796 |
| Jogos | 986 |
| Navegadores | 823 |
| Programação | 944 |
| Rede sociais | 816 |
| Aplicativos de segurança | 1097 |
| Smartphone | 1127 |
| SO | 1183 |
| *Tablets* | 1555 |
| Aplicativos utilitários | 996 |

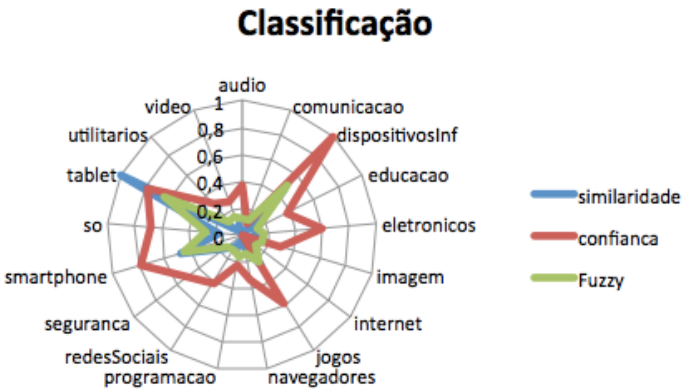
Fonte: Elaborada pela autora

Esse conjunto de dados (Tabela 13) foi projetado sobre a base de termos definindo os vetores para cada categoria.

## 5.2 ANÁLISE DOS RESULTADOS

As figuras e tabelas a seguir ilustram os resultados desse modelo de organização de conteúdo, quando aplicado a quatro documentos testes considerando as 17 categorias. Os resultados apresentados sintetizam os valores de similaridade e confiança apresentados no terceiro capítulo e a aplicação da modelagem *fuzzy*, apresentada no quarto capítulo, em função da saída desfuzzyficada.

Figura 39 – Resultado do classificador para o texto sobre *tablets*



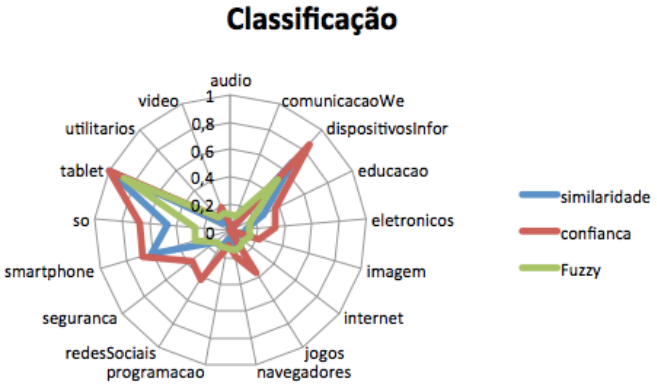

Fonte: Elaborada pela autora

Tabela 14 – Resultado detalhado da classificação do texto sobre *tablets*

| **Categoria** | **Sim.** | **Conf.** | **Saída *fuzzy*** |
|---|---|---|---|
| Áudio | 0,00 | 0,37 | 0,129 |
| Comunicação *web* | 0,00 | 0,10 | 0,123 |
| Dispositivos Informática | 0,46 | 1,00 | 0,500 |
| Educação | 0,09 | 0,37 | 0,129 |
| Eletrônicos | 0,12 | 0,59 | 0,164 |
| Imagem | 0,08 | 0,29 | 0,151 |
| Internet | 0,00 | 0,00 | 0,117 |
| Jogos | 0,25 | 0,59 | 0,233 |
| Navegadores | 0,08 | 0,34 | 0,136 |
| Programação | 0,04 | 0,22 | 0,154 |
| Redes Sociais | 0,09 | 0,41 | 0,122 |
| Segurança | 0,15 | 0,51 | 0,142 |
| Smartphone | 0,47 | 0,78 | 0,450 |
| SO | 0,17 | 0,68 | 0,277 |
| *Tablets* | 1,00 | 0,78 | 0,644 |
| Utilitários | 0,05 | 0,32 | 0,141 |
| Vídeo | 0,08 | 0,27 | 0,157 |

Fonte: Elaborada pela autora

Figura 40 – Resultado do classificador para o texto sobre dispositivos de informática



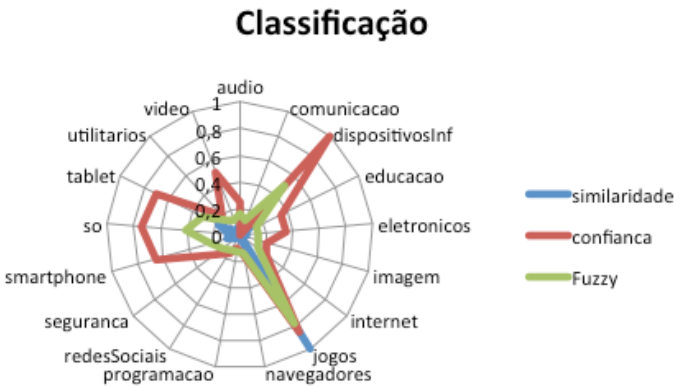

Fonte: Elaborada pela autora

Tabela 15 – Resultado detalhado da classificação do texto sobre dispositivos de informática

| **Categoria** | **Sim.** | **Conf.** | **Saída *fuzzy*** |
|---|---|---|---|
| Áudio | 0,01 | 0,06 | 0,118 |
| Comunicação *web* | 0,00 | 0,00 | 0,117 |
| Dispositivos Informática | 0,68 | 0,86 | 0,512 |
| Educação | 0,27 | 0,36 | 0,158 |
| Eletrônicos | 0,11 | 0,33 | 0,139 |
| Imagem | 0,10 | 0,22 | 0,154 |
| Internet | 0,01 | 0,00 | 0,117 |
| Jogos | 0,16 | 0,36 | 0,136 |
| Navegadores | 0,07 | 0,19 | 0,144 |
| Programação | 0,05 | 0,11 | 0,125 |
| Redes Sociais | 0,13 | 0,42 | 0,129 |
| Segurança | 0,14 | 0,36 | 0,131 |
| *Smartphone* | 0,61 | 0,67 | 0,267 |
| SO | 0,47 | 0,67 | 0,258 |
| *Tablets* | 1,00 | 1,00 | 0,877 |
| Utilitários | 0,08 | 0,14 | 0,131 |
| Vídeo | 0,15 | 0,19 | 0,144 |

Fonte: Elaborada pela autora

Figura 41 – Resultado do classificador para o texto sobre jogos 1



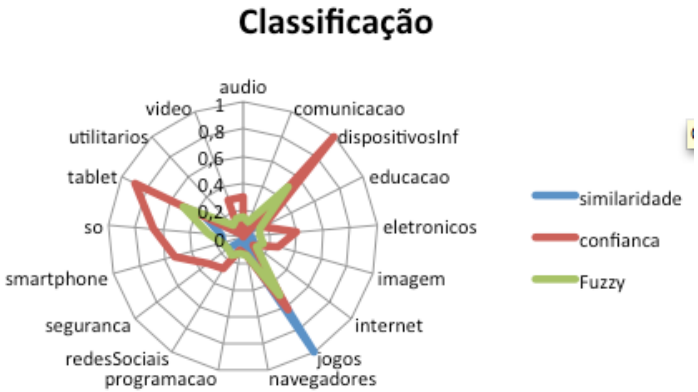

Fonte: Elaborada pela autora

Tabela 16 – Resultado detalhado da classificação do texto sobre jogos 1

| **Categoria** | **Sim.** | **Conf.** | **Saída *fuzzy*** |
|---|---|---|---|
| Áudio | 0,00 | 0,25 | 0,164 |
| Comunicação *web* | 0,00 | 0,00 | 0,117 |
| Dispositivos informática | 0,23 | 1,00 | 0,500 |
| Educação | 0,06 | 0,35 | 0,133 |
| Eletrônicos | 0,05 | 0,35 | 0,133 |
| Imagem | 0,03 | 0,20 | 0,147 |
| Internet | 0,00 | 0,25 | 0,164 |
| Jogos | 1,00 | 0,85 | 0,775 |
| Navegadores | 0,02 | 0,15 | 0,133 |
| Programação | 0,01 | 0,10 | 0,123 |
| Redes Sociais | 0,02 | 0,15 | 0,133 |
| Segurança | 0,04 | 0,25 | 0,164 |
| *Smartphone* | 0,10 | 0,65 | 0,225 |
| SO | 0,08 | 0,75 | 0,410 |
| *Tablets* | 0,18 | 0,70 | 0,316 |
| Utilitários | 0,03 | 0,20 | 0,147 |
| Vídeo | 0,06 | 0,50 | 0,124 |

Fonte: Elaborada pela autora

Figura 42 – Resultado do classificador para o texto sobre jogos 2




Fonte: Elaborada pela autora

Tabela 17 – Resultado detalhado da classificação do texto sobre jogos 2

| **Categoria** | **Sim.** | **Conf.** | **Saída *fuzzy*** |
|---|---|---|---|
| Áudio | 0,00 | 0,30 | 0,147 |
| Comunicação *web* | 0,00 | 0,00 | 0,117 |
| Dispositivos Informática | 0,25 | 1,00 | 0,500 |
| Educação | 0,03 | 0,17 | 0,138 |
| Eletrônicos | 0,06 | 0,40 | 0,123 |
| Imagem | 0,08 | 0,27 | 0,157 |
| Internet | 0,00 | 0,10 | 0,123 |
| Jogos | 1,00 | 0,63 | 0,506 |
| Navegadores | 0,10 | 0,13 | 0,129 |
| Programação | 0,02 | 0,10 | 0,123 |
| Redes Sociais | 0,04 | 0,27 | 0,157 |
| Segurança | 0,14 | 0,33 | 0,139 |
| *Smartphone* | 0,16 | 0,53 | 0,150 |
| SO | 0,16 | 0,67 | 0,258 |
| *Tablets* | 0,30 | 0,90 | 0,497 |
| Utilitários | 0,06 | 0,03 | 0,118 |
| Vídeo | 0,08 | 0,30 | 0,147 |

Fonte: Elaborada pela autora

Nos resultados gerais da avaliação do modelo apresentados nas Figuras 39 a 42 observa-se que esta proposta considerou sempre as três categorias com maior grau de pertinência para cada documento avaliado.

Dessa forma, a avaliação do modelo considerou uma classificação dos documentos avaliados em múltiplas categorias.

Nessa avaliação é possível verificar que normalmente a categoria original do documento, isto é, na qual o documento está previamente indexado, é uma das categorias que o modelo também considera como mais relevante. Além disso, os resultados apresentados pelo modelo são similares aos resultados que seriam obtidos por humanos, ou seja, as demais categorias que o modelo considera relevante também são visivelmente identificadas como relacionadas e pertinentes ao texto. Essa análise ficou clara no texto apresentado pela imagem da Figura 34.

O uso da modelagem *fuzzy* permite especificar um valor que atenua os resultados das variáveis similaridade e confiança; ou seja, a saída desfuzzyficada obtida é um valor que fica entre os resultados das variáveis confiança e similaridade. Desse modo, a utilização da lógica *fuzzy* permite tratar a incerteza presente nos resultados gerados pelas variáveis de similaridade e confiança, além de considerar a sobreposição das categorias que podem ser relacionadas com um mesmo documento.

O modelo desenvolvido permite adquirir conhecimento sobre documentos de interesse para a organização, possibilitando que valores *fuzzy* sejam atribuídos ao documento por meio de uma coleção de categorias de interesse identificadas na organização.

# 6 CONSIDERAÇÕES E RECOMENDAÇÕES

O desenvolvimento do modelo de organização de documentos no contexto de uma memória organizacional considerou a estrutura de uma MO implementada por ontologias. Especificamente uma ontologia de domínio apoia o modelo desenvolvido porque é nessa ontologia que se encontram todas as especificações e detalhes dos conteúdos de interesse para a organização. A ontologia de informação, também citada no modelo, armazena os tipos de documentos: artigos, livros e relatórios da organização.

## 6.1 CONCLUSÕES

Um método para classificação *fuzzy* em múltiplas categorias apoia o modelo desenvolvido. Assim, com base na literatura, definiram-se variáveis que expressam a capacidade de analisar a similaridade e confiança de um documento por meio de uma base de termos, a qual se constrói com uma coleção de documentos já classificados em categorias de interesse para a organização.

Um grupo de resultados sobre documentos processados, em função da similaridade e confiança em relação a uma base de termos da organização, compõe um conjunto de treinamento também denominado de base de conhecimento. Dessa base foi possível identificar um padrão, que definiu um conjunto de regras, por meio de um processo de descoberta de conhecimento. Esse processo envolveu especificamente a mineração dos dados da base de conhecimento. Dessa forma, foi possível definir um modelo geral que é utilizado na definição das regras e funções de pertinência da modelagem difusa para a classificação dos documentos em múltiplas categorias.

O modelo geral das regras identificadas na mineração e implementadas na modelagem *fuzzy* considera quais variáveis são mais significativas e também contribui na especificação das funções de pertinência, como a definição dos termos linguísticos dos conjuntos difusos. As ADs geradas possibilitaram uma visualização mais clara sobre as regras e funções dessa modelagem difusa.

O uso de algoritmos de aprendizagem de máquina para descoberta de conhecimento através da mineração dos dados permite refinar a implementação dos classificadores, porque eles adquirem conhecimento baseado em informações que extraem por meio de

resultados iniciais. Assim, as regras geradas pela descoberta de conhecimento garantem a completude, porque provêm de um conjunto de treinamento que classifica um grande número de exemplos. A consistência do modelo é garantida porque ele tem um valor de acurácia satisfatório.

Um modelo de organização de documentos que tente responder a qual categoria um texto pertence deve considerar com qual grau de pertencimento essa afirmação pode ser dada. O próprio modelo que realiza esse procedimento deve considerar que um documento, muitas vezes, não pertence a somente uma categoria. Essa possibilidade de sobreposições de categorias dá margem para que o problema seja mais bem caracterizado com a abordagem difusa. Ou seja, existe imprecisão na definição da categoria à qual um documento pertence.

A inserção da lógica *fuzzy* no método de classificação permite que sejam estabelecidos procedimentos que considerem a imprecisão das variáveis de entrada da modelagem: a similaridade e a confiança. Além do mais, a modelagem *fuzzy* permite atenuar os resultados das variáveis similaridade e confiança através de uma saída *fuzzy*.

A motivação para implementação desse modelo de organização foi garantir que o método de classificação proposto funcionasse de acordo com seu propósito. Assim, percebeu-se que o modelo representa adequadamente o contexto e as inferências das variáveis mais relevantes para a classificação em uma organização. Além disso, as saídas geradas pelo modelo indicam a relevância do documento para cada uma das categorias consideradas pela organização. As abordagens difusas representam bem os problemas de classificação de documentos, uma vez que, além de ativarem mais de uma função de pertinência, tratam a imprecisão envolvida nesse processo.

O modelo de organização de documentos desenvolvido é genérico, não especializado para uma determinada organização. Assim, é possível trata-lo e utilizá-lo em diferentes contextos e cenários, de acordo com o objetivo ou critério de classificação da organização. Além disso, o uso do modelo não oferece restrições quanto a sua aplicação em diferentes idiomas latinos.

## 6.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

O modelo de organização de documentos desenvolvido tem uma grande importância porque lida com um processo de classificação de documentos. Os processos de classificação de informações textuais se

relacionam diretamente com a quantidade de informações presentes na *web*. Problemas específicos de categorização e recuperação da informação tornam-se cada vez mais importantes, considerando a rápida proliferação de informação textual disponível na internet. Portanto, trabalhos que estudem formas de processar o modelo de classificação em arquiteturas de sistemas distribuídos podem trazer ganhos e contribuições significativos. Dessa forma, é interessante a análise do processo de classificação de informações em bases de dados distribuídas.

Espera-se que a identificação das categorias para um determinado documento seja tal que não conduza a perdas de qualquer tipo de informação relacionada ao contexto do documento. Quanto mais o método de classificação proposto for aplicado, mais dados para análise do modelo são gerados para diferentes categorias. Então, para qualquer organização, é preciso predefinir o conjunto de categorias para determinar o campo de interesse da aplicação dos resultados dessa classificação. O conceito de correspondência entre os documentos analisados e as categorias predefinidas deve ser avaliado para garantir que o modelo proposto esteja atendendo as suas especificações.

Dessa forma, uma proposta que sempre trate do reconhecimento de padrões, em um conjunto de dados gerados para uma base de conhecimento, poderia gerar um modelo com uma acurácia maior, pois consideraria possíveis variações das regras. Assim, o conjunto de regras da modelagem *fuzzy* seria implementado dinamicamente toda vez que uma base de termos fosse gerada.

Além disso, as funções de pertinência poderiam, também, ser implementadas dinamicamente com a utilização de técnicas de algoritmos genéticos ou redes bayesianas. Para isso, os dados do modelo descoberto pela AD, do método de classificação proposto, poderiam servir de parâmetros para alguma dessas técnicas. O uso de funções de pertinência desenvolvidas dinamicamente poderia atenuar ainda mais o resultado da saída entre as variáveis similaridade e confiança.

# REFERÊNCIAS

ABECKER, A.; BERNARDI, A.; HINKELMANN, K.; KUHN, O.; SINTEK, M. Toward a technology for organizational memories. **Intelligent Systems and their Applications**, IEEE, v. 3, p. 40-48, 1998.

ABDUL-RAHMAN, Shuzlina; MUTALIB, Sofianita; KHANAFI, NurAmira; Ali, AzlizaMohd. Exploring feature selection and support vector machine in text categorization. In: COMPUTATIONAL SCIENCE AND ENGINEERING (CSE), 16., 2013, Sydney. **Proceedings…** IEEE, 2013. p. 1101-1104.

ALE, M. A.; GERARDUZZI, C.; CHIOTTI, O.; GALLI, M. R. Organizational knowledge sources integration through an ontology-based approach: the Onto-DOM architecture. **Emerging Technologies and Information Systems for the Knowledge Society,** Springer Berlin Heidelberg, p. 441-450, 2008.

ANTONIE, M. L.; ZAIANE, O. R. Text document categorization by term association., In: IEEE INTERNATIONAL CONFERENCE IN DATA MINING, 2002, Maebashi City. **Proceedings…** IEEE, 2002. p. 19-26.

ARGOTE, Linda; MIRON-SPEKTOR, Ella. Organizational learning: From experience to knowledge. In: **Organization Science**, v. 22, n. 5, p. 1123-1137, 2011.

ARGOTE, Linda. **Organizational learning**: creating, retaining an transferring knowledge. New York: Springer, 2013. p. 85-113,

AZAM, Nouman; YAO, JingTao. Comparison of term frequency and document frequency based feature selection metrics in text categorization. **Expert Systems with Applications**, v. 39, n. 5, 2012. p. 4760-4768.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação:** conceitos e tecnologia das máquinas de busca. 2. ed. São Paulo: Pearson Education Limited, 2011.

BALAMURUGAN, Appavu alias; RAJARAM, Ramasamy; PRAMALA, S; RAJALAKSHMI, S; JEYENDRAN, C; PRAKASH, J. Dinesh Surya. NB+: An improved Naïve Bayesian algorithm. **Knowledge-Based Systems**, v. 24, n. 5, p. 563-569, 2011.

BANG, S. L.; YANG, J. D.; YANG, H. J. Hierarchical document categorization with k-NN and concept-based thesauri. **Information Processing and Management**, v. 42, n. 2, p. 387-406, 2006.

BARTH, Fabrício Jailson. **Recuperação de documentos e pessoas em ambientes empresariais através de árvores de decisão.** 2009. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica da Universidade de São Paulo, São Paulo, 2009.

BORDOGNA, Gloria; PASI, Gabriella. **Personalised indexing and retrieval of heterogeneous structured documents**. Information Retrieval, v.8, n. 2, apr. 2005.

BRAGA, Luis Paulo Vieira. **Introdução à mineração de dados**. 2. ed. ampl. e rev. Rio de Janeiro: Editora E-papers, 2005.

CLIFTON, Christopher. **Data mining**. Encyclopædia Britannica. Disponível em: <http://www.britannica.com/EBchecked/topic/1056150/data-mining>. Acesso em: 15 mar. 2013.

CHERMAN, Everton Alvares; MONARD, Maria Carolina; METZ, Jean. Multi-label problem transformation methods: a case study. **CLEI Electronic Journal**, v. 14, n. 1, p. 4, 2011.

COLACE, Francesco; SANTO, Massimo de; GRECO, Luca; NAPOLETANO, Paolo. Text classification using a few labeled examples. **Computers in Human Behavior**, v. 30, p. 689-697, 2014.

COUSSEMENT, K.; VAN DEN POEL, D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. **Information Management**, v. 45, n. 3, p. 164-174, 2008.

DAN, Li; LIHUA, Liu; ZHAOXIN, Zhang. Research of text categorization on WEKA. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEM DESIGN AND ENGINEERING

APPLICATIONS (ISDEA), 3., Hong Kong. **Abstracts…** ISDEA, 2013. p. 1129-1131.

DOW, Kevin E.; HACKBARTH, Gary; WONG, Jeffrey. Data architectures for an organizational memory information system. **Journal of the American Society for Information Science and Technology,** v. 64, n. 7, p. 1345-1356, 2013.

DZIEKANIAK, Gisele. Tecnologias de descoberta de conhecimento na gestão do conhecimento: contextualizações com a sociedade do conhecimento. **Data Grama Zero**: revista de Ciência da Informação, v. 11, n. 1, fev. 2010.

FRANÇA, J. B. S.; SANTORO, F. M.; BAIÃO, F. A. Towards characterizing knowledge intensive processes. In**:** INTERNATIONAL CONFERENCE ON COMPUTER-SUPPORTED COOPERATIVE WORK IN DESIGN, 16., 2012, Wuhan. **Proceedings…** IEEE, v. 113, p. 497-504, 2012.

FRANÇA, J. B. S.; NETTO, J. M.; CARVALHO, J. E. S.; BAIÃO, F. A.; SANTORO, F. M.; PIMENTEL, M. An exploratory study on collaboratively conceptualizing knowledge intensive processes. In: INTERNATIONAL CONFERENCE ON BUSINESS PROCESS MODELING, DEVELOPMENT AND SUPPORT, 13., Gdansk. **Proceedings… Lectures Notes in Business information Process**, v. 113. Berlin: Springer, 2012. p. 46-60.

GAO, Guanyu; GUAN, Shengxiao. Text categorization based on improved Rocchio algorithm. In: INTERNATIONAL CONFERENCE ON SYSTEMS AND INFORMATICS (ICSAI), 2012, Yantai. **Proceedings…** IEEE, 2012. p. 2247-2250.

GODBOLE, S.; SARAWAGI, S. Discriminative methods for multi-labeled classification. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING – PAKDD, 8., 2004, Sydney. **Proceedings…** Berlin: Springer, 2004. p. 22-30.

GUO, G.; WANG, H.; BELL, D.; BI, Yaxin; GREER, Kieran. kNN model-based approach and its application in text categorization. **Computational Linguistics and Intelligent Text Processing**, LNCS, n. 2945, p. 559-570, 2004.

GUAR Edgard Blücher, 2000.

HAO, Pan; YING, Duan; LONGYUAN, Tan. Application for web text categorization based on support vector machine. In: INTERNATIONAL FORUM ON COMPUTER SCIENCE-TECHNOLOGY AND APPLICATIONS – IFCSTA, 2009, Chongqing. **Proceedings…** Piscataway: IEEE, 2009. p. 42-45.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning:** data mining, inference, and prediction. 2. ed. Berlin: Springer, 2009.

HERRERA-VIEDMA, E. An information retrieval model with ordinal linguistic weighted queries based on two weighting elements. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 9, suppl. p. 77-87, sep. 2001.

HRISTEA, Florentina T. **The Naïve Bayes model for unsupervised word sense disambiguation**. Berlin: Springer, 2013.

IKONOMAKIS, M.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification using machine learning techniques. **WSEAS Transactions on Computers**, v. 4, n. 8, p. 966-974, 2005.

INTEXT MINING. Disponível em: <http://www.intext.com.br/>. Acesso em: 16 jun. 2013.

JACKSON, P. **An exploratory survey of the structure and components of organizational memory**. Heidelberg: Physica-Verlag HD, 2008. p. 89-110.

JIANG, Shengyi; PANG, Guansong; WU, Meiling; KUANG, Limin. **An improved K-nearest-neighbor algorithm for text categorization**. Disponível em: < http://www.researchgate.net/publication/232406523_An_improved_K-nearest-neighbor_algorithm_for_text_categorization >. Acesso em: 16 jun. 2013.

JIANG, Jung-Yi; TSAI, Shian-Chi; LEE, Shie-Jue. FSKNN: Multi-label text categorization based on fuzzy similarity and k nearest neighbors.

**Expert Systems with Applications:** an international journal, v. 39, n. 3, p. 2813-2821, 2012.

JONES, K. S.; WALKER, S.; ROBERTSON, S. A probabilistic model of information retrieval: development and comparative experiments. **Information Processing and Management**, v. 36, n. 6, p. 779-808, 2000.

KACPRZYK, J.; PASI, G.; VOJTÁS, P.; ZADROZNY, S. Fuzzy querying: issues and perspectives. **Kybernetika**, v. 6, n. 36, p. 605-616, 2000.

KARABULUT, Mustafa. Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection. **Knowledge-Based Systems**, v. 54, p. 288-297, 2013.

KIM, Tai-hoon; YANG, Laurence Tianruo; PARK, Jong Hyuk; VASILAKOS, Thanos; YEO, Sang-Soo. In: YANG, L.T., PARK**,** J.H., VASILAKOS**,** T., YEO**,** S.-S. (Ed.). **Advances in Communication and Networking:** Second International Conference on Future Generation Communication and Networking, FGCN 2008, Sanya, Hainan Island, China, December 13-15, 2008. Berlin: Springer, 2009.

KUMAR, R. Dinesh; SUGANITHI, J. Offline handwritten sanskrit character recognition using support vector machines. **Journal of Environmental Science**, v. 2, n. 3, p. 769-775, 2013.

LEE, Kyung-Soon; KAGEURA, Kyo. A. Virtual relevant documents in text categorization with support vector machines. **Information Processing & Management**, v. 43, n. 4, p. 902-913, 2007.

LOMAX, Susan; VADERA, Sunil. A survey of cost-sensitive decision tree induction algorithms. **ACM Computing Surveys (CSUR)**, v. 45, n. 2, p. 16, 2013.

LUO, X.; ZINCIR-HEYWOOD, A. N. Evaluation of two systems on multi-class multi-label document classification. In: INTERNATIONAL SYMPOSIUM ON METHODOLOGIES FOR INTELLIGENT SYSTEMS, 15., 2005, Saratoga Springs. **Proceedings…** Berlin: Springer, 2005.

MAAZOUZI, Faiz; BAHI, Halima. Using multi decision tree technique to improving decision tree classifier. **Int. J. Business Intelligence and Data Mining**, v. 7, n. 4, p. 274-287, 2012.

MANNE, Suneetha; FATIMA, S. Sameen. An extensive empirical study of feature terms selection for text summarization and categorization. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE, ENGINEERING AND INFORMATION TECHNOLOGY, (CCSEIT '12), 2., Coimbatore. **Proceedings…** New York: ACM, 2012.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MENG, Wang; LANFEN, Lin; JING, Wang; PENGHUA, Yu; JIAOLONG, Liu; FEI, Xie. Improving short text classification using public search engines. **Integrated uncertainty in knowledge modelling and decision making.** Berlin: Springer, 2013. p. 157-166.

MING, Li; LU, Liu; CHUAN-BO, Li. An approach to expert recommendation based on *fuzzy* linguistic method and *fuzzy* text classification in knowledge management systems. **Expert Systems with Applications**: an international journal, v. 38, n. 7, p. 8586-8596, 2011.

MIRANDA DA SILVA, Heide. Gestão do conhecimento e inteligência competitiva em organizações: uma abordagem conceitual. **Revista de Iniciação Científica da FFC**, v. 7, n. 1, 2008.

MODI, Hiteshri; PANCHAL, Mahesh. Experimental comparison of different problem transformation methods for multi-label classification using MEKA. **International Journal of Computer Applications**, v. 59, n. 15, p. 10-15, 2012.

MOSCHITTI, Alessandro; JU, Qi; JOHANSSON, Richard. Modeling topic dependencies in hierarchical text categorization. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 50. **Long Papers**, v. 1, p. 759-767, 2012.

NEUMANN, Clóvis. **Gestão de sistemas de produção e operações**: produtividade, lucratividade e competitividade. Rio de Janeiro: Elsevier, 2013. 304 p.

NONAKA, Ikujiro; TOYAMA, Ryoko; NAGATA, Akiya. A firm as a knowledge-creating entity: a new perspective on the theory of the firm. **Industrial and corporate change**, v, 9, n. 1, p. 1-20, 2000.

PANG, Guansong; JIANG, Shengyi. A generalized cluster centroid based classifier for text categorization. **Information Processing & Management**, v. 49, n. 2, p. 576-586, 2013.

PAPPA, Gisele L; FREITAS, Alex. **Automating the design of data mining algorithms:** an evolutionary computation approach. Berlin: Springer, 2009.

QI, X.; DAVISON, B.D. Web page classification: features and algorithms. **ACM Computing Surveys**, v. 41, n. 2, 2009.

RAPIDMINER. Disponível em: <http://rapid-i.com/>. Acesso em: 22 mar. 2014.

READ, J. A pruned problem transformation method for multi-label classification. **New Zealand Computer Science Research Student Conference** (NZCSRS 2008), p. 143-150, 2008.

REN, Fuji; SOHRAB, Mohammad Golam. Class-indexing-based term weighting for automatic text classification. **Information Sciences**, v. 236, p. 109-125, 2013.

RUSSELL, Stuart Jonathan; NORVIG, Peter. **Artificial intelligence:** a modern approach. 3. ed. London: Prentice-Hall, 2009.

SASIETA, Héctor Andrés Melgar; BEPPLER, Fabiano Duarte; PACHECO, Roberto Carlos dos Santos. A memória organizacional no contexto da engenharia do conhecimento. **Revista Data Grama Zero**, v. 12, n. 3, 2011.

SATO, Paulo. É possível calcular quantas palavras surgem por dia na Língua Portuguesa? **Revista Nova Escola**, 2009. Disponível em: <http://revistaescola.abril.com.br/lingua-portuguesa/fundamentos/possivel-calcular-quantas-palavras-surgem-dia-lingua-portuguesa-473887.shtml>. Acesso em: 20 mar. 2014.

SESTATNET. **Ensino-aprendizagem de estatística na web**. Disponível em: <http://www.sestatnet.ufsc.br/>. Acesso em: 20 maio 2014.

SHAREF, Nurfadhlina Mohd; KASMIRAN, Khairul Azhar. Examining text categorization methods for incidents analysis. **Intelligence and Security Informatics**, p. 154-161, 2012.

SUPYUENYONG, Varintorn; SWIERCZEK, Fredric William. Knowledge management process and organizational performance in SMEs. **International Journal of Knowledge Management (IJKM)**, v. 7, p. 1-21, 2011.

TAN, S. An effective refinement strategy for KNN text classifier. **Expert Systems with Applications**, v. 30, n. 2, p. 290-298, 2006.

TORRES-NIÑO, Javier; RODRÍGUEZ-GONZÁLEZ, Alejandro; COLOMO-PALACIOS, Ricardo; JIMÉNEZ-DOMINGO, Enrique; ALOR-HERNANDEZ, Giner. Improving accuracy of decision trees using clustering techniques. **Journal of Universal Computer Science**, v. 19, n. 4, p. 484-501, 2013.

TSOUMAKAS, Grigorios; KATAKIS, Ioannis. Multi-label classification: an overview. **International Journal of Data Warehousing and Mining**, v.3, n. 3, p. 1-13, 2007.

TSOUMAKAS, Grigorios; KATAKIS, Ioannis; VLAHAVAS, Ioannis. **Mining multi-label data**. Data mining and knowledge discovery handbook. Berlin: Springer, 2010. p. 667-685.

TOLEDO, Carlos M; ALE, Mariel A; CHIOTTI, Omar; GALLI, María R. An ontology-driven document retrieval strategy for organizational knowledge management systems. **Electronic Notes in Theoretical Computer Science**, v. 281, p. 21-34, 2011.

UYSAL, AlperKursat; GUNAL, Serkan. The impact of preprocessing on text classification. **Information Processing and Management**: an international journal, v. 50, n. 1, jan. 2014.

VINOT, Romain; YVON, François. Improving Rocchio with weakly supervised clustering. **Machine Learning**: ECML.Berlin: Springer, 2003. p. 456-467.

WEI, Chih-Ping; YANG, Chin-Sheng; LEE, Ching-Hsien; SHI, Huihua; YANG, Christopher C. Exploiting poly-lingual documents for improving text categorization effectiveness. **Decision Support Systems**, v. 57, p. 64-76, jan. 2014.

WILGES, B., MATEUS, G., BASTOS, R.; DANTAS M. A case-comparison study of automatic document classification utilizing both serial and parallel approaches. **Journal of Physics**: Conference Series, High Performance Computing Symposium (HPCS)**,** 2013.

YANG, Jieming; LIU, Yuanning; ZHU, Xiaodong; LIU, Zhen; ZHANG, Xiaoxu. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. **Information Processing & Management**, v. 48, n. 4, p. 741-754, 2012.

ZADROŻNY, Slawomir; KACPRZYK, Janusz. Computing with words for text processing: an approach to the text categorization. **Information Sciences**, n. 176, p. 415-437, 2006.

ZHANG, M. L.; ZHOU, Z. H. ML-KNN: A lazy learning approach to multi-label learning. **Pattern Recognition**, v. 40, n. 7, p. 2038-2048, 2007.

ZHOU, B.; YAO, Y. Y.; LUO, J. A three-way decision approach to email spam filtering. IN: CANADIAN CONFERENCE ON ARTIFICIAL INTELLIGENCE (CANADAIN AI'10), 23., Quebec. **Lecture notes in computer science**, v. 6085, p. 28-39, 2010.